Speech Emotion Recognition using LSTM and RNN

R. Nihaal Datta Sai
B.Tech, Computer Science
Dr. M.G.R Educational
And Research Institute
Chennai, India
nihaalsai@gmail.com

Syed Shahbaaz
B.Tech, Computer Science
Dr. M.G.R Educational
And Research Institute
Chennai, India
syedshahbaaz@gmail.com

U.Bhanu Prakash
B.Tech,Computer Science
Dr. M.G.R Educational
And Research Institute
Chennai,India
bhanu.udhissa@gmail.com

ABSTRACT

Deep learning methods are being applied in various recognition tasks such as image, speech, and music recognition. Convolutional Neural Networks (CNNs) especially show remarkable recognition performance for computer vision tasks. In addition, Recurrent Neural Networks (RNNs) and Long short-term memory (LSTM) show considerable success in many sequential data processing tasks. In this study, we investigate the result of the Speech Emotion Recognition (SER) algorithm based on LSTM-RNN trained using an emotional speech database. The main goal of our work is to propose a SER method based on concatenated CNNs and . By applying the propose methods to an emotional speech database, the classification result was verified to have better accuracy than that achieved using conventional classification methods. As Emotion recognition from speech signals is an important but challenging component of Human Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. This paper presents an overview of Deep Learning techniques and discusses some recent literature where these methods are utilized or speech-based emotion recognition. The review covers databases used, emotions extracted contributions made toward speech emotion recognition and limitations related to it.

INTRODUCTION

Speech Emotion Recognition is one of the most challenging tasks in speech signal analysis domain. The importance of emotion recognition is getting popular with improving user experience and with the engagement of Voice User Interfaces. It's an easy task for humans to understand the emotions, but a bit harder for systems to understand. Multimedia pattern recognition is an emerging technology that can extract and analyse large amounts of multimedia information from video and audio sources. In recent years, there has been a drastic growth in the application of machine learning technology using deep learning to solve various recognition problems. Speech Emotion Recognition (SER) is an especially significant task in understanding the characteristics of speech in media. However, recognizing emotions from speech is a very challenging problem because people express emotions in different ways, and the features are unclear to distinguish the emotions. Actually, the paralinguistic problem is challenging even for humans. Conventional techniques for solving this problem are extracting low-level descriptors and training the machine appropriately through learning those features. These methods have been accepted as state of the art for many years in machine learning. However, selecting good features to extract is difficult, and optimization is even more difficult, often being significantly time-consuming in research, development, and validation. Because of this, the traditional trend in speech/audio information retrieval is to focus on the use of powerful strategies for semantic analysis, often relying on model selection to optimize the results. However, deep neural architectures can share low-level representations and naturally progress from low-level to high level structures. The typical method for analysing the audio and speech signals is 2D representation. Time frequency analysis is commonly used in audio processing. We transform the speech signal to 2D representation using Short Time Fourier Transform (STFT) after pre-processing. Then the 2D representation is analysed through CNNs and Long Short-term Memory (LSTM) architectures. Deep learning involves hierarchical representations with increasing levels of abstraction. By traversing sequentially constructed networks, the results corresponding to each selected

audio frame are classified using a sum of probabilities. Recurrent neural networks (RNNs) have been widely used in sequence learning problems such as action recognition, scene labelling and language processing, and have achieved impressive results. Compared with the feed-forward networks such as the convolutional neural network (CNNs), a RNN has a recurrent connection where the last hidden state is an input to the next state. Long Short-Term Memory Network is an advanced RNN, a sequential network, that allow information to persist. It is capable of handling the vanishing gradient problem faced by RNN. A recurrent neural network is also known as RNN is used for persistent memory. Let us say while watching a video you remember the previous scene or while reading a book you know what happened in the earlier chapter. Similarly, RNNs work, they remember the previous information and use it for processing the current input. The shortcoming of RNN is, they cannot remember Long term dependencies due to vanishing gradient. LSTMs are explicitly designed to avoid long-term dependency problems.

LITERATURE SURVEY

NATIONAL LEVEL STATUS Yamuna Devi C Ra, Saishiva Ka, Sunil Kumara, S H Manjulaa, K R Venugopala, L M Patnaik have published the paper about "Comparative study of CNN and RNN for Natural Language Processing". It is published in the year 2018. This paper presents Convolutional Neural Networks and Recurrent Neural Networks, the two main types of DNN architectures, are widely explored to handle various NLP tasks. CNN is supposed to be good at extracting position invariant features and RNN at modeling units in sequence. This work is the first schematic comparison of CNN and RNN on a wide range of representative NLP tasks. The merits and Demerits of this are that they show improvements of about 8% relative in perplexity over standard recurrent neural network LMs. This paper fails to analyze the differences between standard and LSTM networks and the impact on the recognition quality of a speech recognizer. Mr. Santhana Krishnan.J, Dr. Geetha.S have published the paper "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks". It is published in the year 2017. This paper presents about how the Convolutional Neural Networks (CNNs) especially show remarkable recognition performance for computer vision tasks. In addition, Recurrent Neural Networks (RNNs) show considerable success in many sequential data processing tasks. proposed a SER method based on concatenated CNNs and RNNs without using any traditional hand-crafted features. By applying the proposed methods to an emotional speech database, the classification result was verified to have better accuracy than that achieved using conventional classification methods. This paper fails to study the audio/video based multimodal emotion recognition task. Ruhul amin khali, Edward Jones, Mohammad inayatullah Babar, Thamer Alhussain have published the paper "Speech Emotion Recognition Using Deep Learning Techniques". It is published in the year 2019. This paper presents that Convolutional neural network are better in image data analysis, efficient way to use Long-short-term memories in RNN for audio and video classification. Feed-Forward networks only exploit a fixed context length to predict the next output of a sequence, conceptually, standard recurrent networks are difficult to train and therefore are unlikely to show the full potential of recurrent models. These are addressed by a Long Short Term Memory network architecture. The Merits and Demerits of this are the weighted accuracy of the proposed emotion recognition system is improved up to 12% compared to the CNN based emotion recognition system

INTERNATIONAL LEVEL STATUS

Shuai Li, Wanquing Li, Chris Cook, Ce Zhu, Yanbo Gao have published the paper "Independently Recurrent Neural Network (IndRNN): Building a Longer and Deeper RNN". It is published in the year 2017. This paper presents that RNN's are difficult to train due to the well-known gradient vanishing and exploding problems independently recurrent neural networks is proposed in this paper, where neurons in the same layer are independent of each other and they are connected across the layers. Moreover, an IndRNN can work with non-saturated activation functions such as relu and be still trained robustly. The Merits and Demerits are known as gradient vanishing problem in DNN is solved using IndRNN, where neurons in same layer are independent of each other, we can also solve gradient vanishing problem by implementing Long-Short-termmemories(LSTM) in RNN. Jinkyu Lee and Ivan Tashev have published the paper "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition". It is published in the year 2019. This paper talks about the proposed system takes into account the long-range context effect and the uncertainty of emotional label expressions. To extract high level representation of emotional states with regard to its temporal dynamics, a powerful learning method with a bidirectional long, short term memory (BLSTM) model is adopted. The Merits and Demerits are that to overcome the uncertainty of emotional labels, such that all frames in the same utterance are mapped into the same emotional label, it is assumed that the label of each frame is regarded as a sequence of random variables. Then the sequences are trained by the proposed learning algorithm. Seyedmahdad Mirsamadi, Emad Barsoum, Cha Zhang have published the paper "Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention". It is published in the year 2017. This paper presents how it is used deep recurrent neural network to

automatically discover emotionally relevant features from speech. It is shown that using RNN we can learn both the short-time frame-level acoustic features that are emotionally relevant. The Merits and Demerits are the preliminary results show that in general adding a pooling layer on top of the LSTM layers produces the better performance, and the weighted pooling with attention model further improves over mean-pooling by about 1-2% on IEMOCAP data set

PROPOSED SYSTEM

The typical method for analyzing the audio and speech signals is 2D representation. Time-frequency analysis is commonly used in audio processing. We transform the speech signal to 2D representation using Short Time Fourier Transform (STFT) after pre-processing. Then the 2D representation is analyzed through CNNs and Long Short-Term Memory (LSTM) architectures. Deep learning involves hierarchical representations with increasing levels of abstraction. By traversing sequentially constructed networks, the results corresponding to each selected audio frame are classified using a sum of probabilities.

RNN (RECURRENT NEURAL NETWORK):

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit a temporal dynamic behavior, which are derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech emotion.

RNNs come in many variants

Fully recurrent neural networks

Fully recurrent neural networks (FRNN) connect the outputs of all neurons to the inputs of all neurons. This is the most general neural network topology because all other topologies can be represented by setting some connection weights to zero to simulate the lack of connections between those neurons.

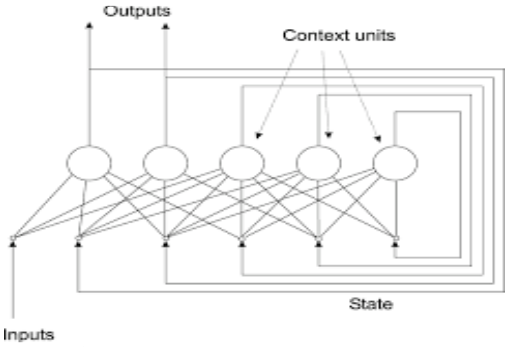

Fig.1.FULLY CONNECTED RNN

Independently RNN

The Independently recurrent neural network addresses the gradient vanishing and exploding problems in the traditional fully connected RNN. Each neuron in one layer only receives its own past state as context information (instead of full connectivity to all other neurons in this layer) and thus neurons are independent of each other's history.

Second order RNNs

Second order RNNs use higher order weights instead of the standard weights, and states can be a product. This allows a direct mapping to a finite state machine both in training, stability, and representation. Long short-term memory is an example of this.

LONG SHORT-TERM MEMORY

Long short-term memory (LSTM) is a deep learning system that avoids the vanishing gradient problem. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). LSTM is normally augmented by recurrent gates called “forget gates”. LSTM prevents backpropagated errors from vanishing or exploding. Instead, errors can flow backwards through unlimited numbers of virtual layers unfolded in space. That is, LSTM can learn tasks that require memories of events that happened thousands or even millions of discrete time steps earlier.

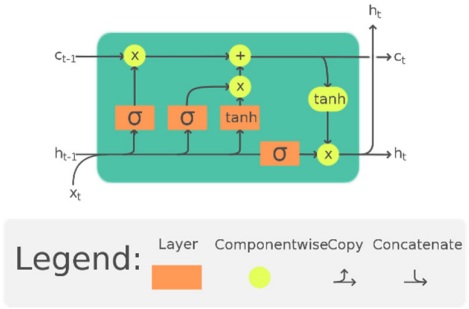

Fig.2. LSTM DIAGRAM

DENSE LAYER

It is the regular deeply connected neural network layer. It is most common and frequently used layer. Dense layer does the below operation on the input and return the output.

Output = activation (dot (input, kernel) + bias)

were,

• input represent the input data • kernel represent the weight data

• dots represent NumPy dot product of all input and its corresponding weights

• bias represent a biased value used in machine learning to optimize the model.

• Activation represents the activation function.

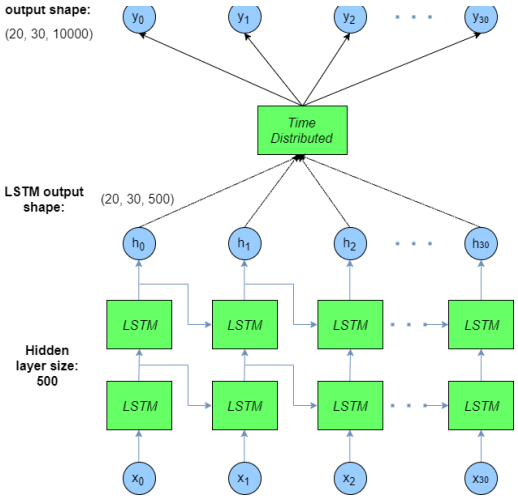

Fig.3. Dense layers along with LSTM layers, Inputs and Outputs

ACTIVATION FUCNTION

We know that an activation is required between matrix multiplications to afford a neural network the ability to model non-linear processes. A classical LSTM cell already contains quite a few non-linearities: three sigmoid functions and one hyperbolic tangent (tanh) function, here shown in a sequential chain of repeating (unrolled) recurrent LSTM cells: So there are plenty of non-linearities being used, meaning a ReLU activation function is used after a fully-connected layer. Specifically, the way ReLU works is it returns input directly if the value is greater than 0. If less than 0, then 0.0 is simply returned.

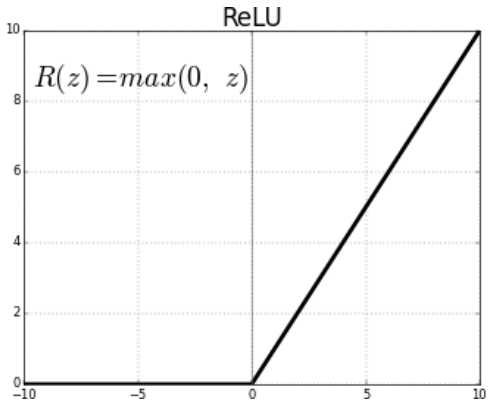

Fig.4. ReLU Activation Function

SOFTMAX

The SoftMax function is a function that turns a vector of K real values into a vector of K real values that sum to 1. The input values can be positive, negative, zero, or greater than one, but the SoftMax transforms them into values between 0 and 1, so that they can be interpreted as probabilities. The output of the SoftMax function is equivalent to a categorical probability distribution, it tells you the probability that any of the classes are true.

LOSS FUNCTION

Cross Entropy

Formally, it is designed to quantify the difference between two probability distributions. The cross entropy between two probability distributions and over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set, if a coding scheme is used that

is optimized for an "unnatural" probability distribution rather than the "true" distribution

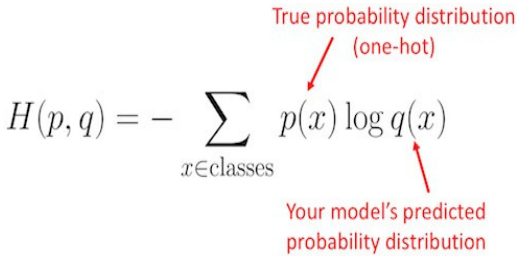

Fig.5. Cross Entropy.

DROUPOUTS

Dropout is a regularization method where input and recurrent connections to LSTM units are probabilistically excluded from activation and weight updates while training a network. This has the effect of reducing overfitting and improving model performance.

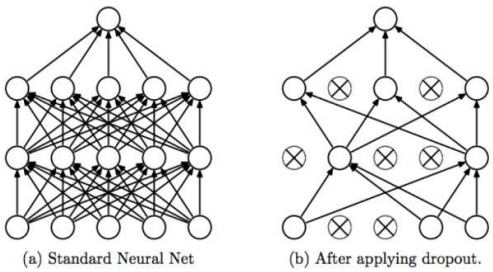

Fig.6. Implementation of Dropout Layer

DATASET

Dataset provides information to the module that can learn the methods and implements the task successful. In this project we used Ryerson audio-visual database for emotional speech and song (RAVDESS)dataset which contains 1440 audio files of 24 acters from which 12 males and 12 females. It contains 8 different emotions which are neutral-01, calm-02, happy-03, sad-04, angry05, fearful-06, disgust-07, surprise-08. The 7th and 8th number will determine the emotion of the audio

CONCLUSION

In this project Speech-Emotion-Recognition using LSTM-RNN has been implemented using Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) as data set. Achieved an accuracy which is greater than 95%, and we have built a web-interface for emotion recognition using model weights.

FUTURE SCOPE

Emotion recognition is most important now a days to improve the relation between human and artificial intelligence, present artificial intelligences like Siri and Alexa can be programmed to play music or make interacts with humans based on the speech, Speech Emotion detection can analyse the mood of the person and we can using IOT and adjust the lighting and music in room to improve mood and can suggest some shows to help users to always be in happy mood, In medical field we can track and monitor the emotional state of the patient and can provide suitable medical treatment, One important goal is to enable computers to understand the emotional states expressed by the human subjects, so that personalized responses can be delivered accordingly.

REFERENCES:

[1] A. Coates, A. Y. Ng, and H. Lee. An analysis of single-layer networks in unsupervised feature learning. In International Conference on Artificial Intelligence and Statistics, pages 215–223, 2011.

[2] I. Goodfellow, D. Warde-farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 1319– 1327, 2013.

[3] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Computer Science Department, University of Toronto, Tech. Rep, 2009.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[5] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus. Regularization of neural networks using dropconnect. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), pages 1058–1066, 2013.