Speech Emotion Recognition Using Convolutional-Recurrent Neural Networks with Attention Model

YAWEI MU, LUIS A. HERNÁNDEZ GÓMEZ, ANTONIO CANO MONTES, CARLOS ALCARAZ MARTÍNEZ, XUETIAN WANG and HONGMIN GAO

ABSTRACT

Speech Emotion Recognition (SER) plays an important role in human-computer interface and assistant technologies. In this paper, a new method is proposed using distributed Convolution Neural Networks (CNN) to automatically learn affect-salient features from raw spectral information, and then applying Bidirectional Recurrent Neural Network (BRNN) to obtain the temporal information from the output of CNN. In the end, an Attention Mechanism is implemented on the output sequence of the BRNN to focus on target emotion-pertinent parts of an utterance. This attention mechanism not only improves the classification accuracy, but also provides model’s interpretability. Experimental results show that this approach can gain 64.08% weighted accuracy and 56.41% unweighted accuracy for four-emotion classification in IEMOCAP dataset, which outperform previous results reported for this dataset.

KEYWORDS

INTRODUCTION

With the development of Artificial Intelligence, more and more studies and research attempt to investigate Speech Emotion Recognition (SER), which not only makes the communication between machine and human more natural and real, but also has great potential in the development of auxiliary technologies, such as clinical assistant to help monitor the patient’s emotional change, especially those who have difficulty to express themselves very well. Speech emotion recognition is a sequence classification problem, where the input is the length-variant utterance and the output is the actual emotion. Considering the characteristics of speech and various ways people express emotions, there are many challenges to overcome. Besides, there is still no consensus about the best features to recognize emotion from speech among academic community or industrial circles. In recent years, there have been many researches involving in this area.
In a general way, it can be stated that there are two main strategies to address the development of SER. The first one is to manually extract as many as relevant short-term features from audio signal, such as MFCC, pitch and energy etc. [1,2]. The second one is to let a neural network extract features automatically, for example, using autoencoders [3,4], or CNN models [5]. W.Q. Zheng et al. [6] used deep convolutional neural networks (DCNN) on five emotions from IEMOCAP [7], achieving 40.02% accuracy in speaker-independent manner.

After affect-salient features are available, several classification algorithms can be applied, such as SVM [8], Hidden Markov Method [9], Random Forest [10], DNN [11], or RNN [10]. Convolutional Recurrent Neural Network (CRNN), a combination of CNN followed by Long Short-Term Memory (LSTM) networks, has been investigated in [12], showing very good performance on Berlin [13] dataset, which is a much smaller database than IEMOCAP.

Based on that previous research, in this paper we propose a similar architecture (CNN + LSTM) extended with an attention mechanism. The proposed model is tested on IEMOCAP.

More specifically, in the proposed architecture a Convolutional Neural Network is firstly applied to the two-dimensional spectral representation of the audio signal to extract emotion-relevant features from short term segments. Then Bidirectional Recurrent Neural Networks (BRNN) are further implemented to model the temporal information closely associated to target emotion. Inspired by [14], an attention model is used to weight the output sequence of BRNN and get the final emotion classification. The experimental results show that the combination of CNN and BRNN (CRNN) as well as the use of an attention mechanism can improve results previously reported on IEMOCAP dataset.

The rest of paper is organized as follows. Details on the use of IEMOCAP are firstly described in section 2. The CRNN model proposed for Speech Emotion Recognition is introduced in section 3. Afterwards, experimental setup and results are presented in section 4. Conclusions and future research are given at the end of the paper.

### EXPERIMENTAL DATASET

To compare and evaluate our proposed method, we use IEMOCAP dataset in our experiments, which is widely-used for speech emotion recognition. It contains audio data from 10 actors (5 females and 5 males). There are five sessions, which consist of dialogues between different pair of male and female actors. Only 4490 among all IEMOCAP utterances were considered reliable and used as they were labeled as the same emotion by the majority of three annotators. Therefore, these utterances constitute all the dataset used for model training and testing as it will be described later. We performed SER classification on four emotions, including anger, happy, neutral and sad. The distribution of these four emotions in each session and all sessions are shown in Fig. 1. Obviously, these four emotions have highly unbalanced proportions. The number of utterances corresponding to neutral emotion is approximately 3 times more than for happy emotions. Therefore, it is necessary to bias the model to pay more attention to the minority class. Many tactics can be chosen to combat this problem, such as oversampling [15], panelizing models with cost sensitive
classifier [16], weighting the cost function to balance the classes [17] and so forth. In this paper, we chose cost-sensitive function for its simplicity.

PROPOSED SER MODEL

As Fig. 2 illustrates, to feed our SER model, audio signals are first represented by time-frequency maps, which are obtained by Fast Fourier transform of the speech wave form. We used 256 FFT bins for each short-term frame, which has the length of 20ms with 10ms overlapping. Subsequently, we stack a context window of 32 adjacent FFT frames together as one mid-term segment, which results in $32 \times 128$ two-dimensional time-frequency array. Thus, the total length of one segment is $10ms \times 32 + 10ms = 330ms$. 50% overlapping between segments was used to make the analysis of neural networks smoother. In general, emotion is carried by continuous segments, the length of which is highly dependent on the speakers and certain emotion. However, a segment longer than 250ms has been verified to contain efficient emotion information [18, 19].

Once input audio signal is represented as spectrogram segments, they share the same property as a sequence of images, from which CNN can be applied to extract hierarchical affect-salient features (see Fig. 2 and Fig. 3). Afterwards, as shown in Fig. 2, bidirectional LSTM can be further used to explore the dynamic temporal information related to the final emotion representation hidden in the sequence of mid-term segments. Before the final classification, attention interface is implemented to focus on those more discriminative part of the RNN output sequence. Fig. 2 shows the detailed schema of our proposed approach.

![Figure 1. Utterance distribution about four emotions.](image-url)
Figure 2. The proposed CRNN with attention model.

Figure 3. The convolutional architecture for one mid-term segment.

Figure 4. Histogram of number of segments in all utterance.
CONVOLUTIONAL NEURAL NETWORKS

In CNN input is usually considered as an image, so some certain properties can be encoded, such as local connectivity and weight sharing, which substantially reduces the number of parameters to be trained in the network. In our case, CNN plays a very important role in extracting discriminative features from the time-frequency representation of audio signal, which can be viewed as a sequence of images. As it can be seen in Fig. 2, the feature representation provided by the CNN layer is used as input to the Bidirectional LSTM layer. LSTM is applied in a parallel and time-distributed manner.

Also, as shown in Fig. 3, for the analysis of each mid-term spectrogram segment, CNN has the basic architecture of [INPUT-CONV-RELU-POOL-CONV-RELU-POOL-FLAT].

The CNN input’s size is 32*128, which is one mid-term segment from the time-frequency map. With 50% overlapping between segments, the distribution of the number of segments in the whole dataset is shown in Fig. 4.

To preserve as much as possible the information provided by the IEMOCAP, we choose the padding length with 78 segments for every utterance sample. Therefore, the complete convolutional net will contain 78 parallel time-distributed architectures as showed in Fig. 2.

BIDIRECTIONAL RECURRENT NEURAL NETWORKS

We assume that in modeling emotions both past and future audio information can be relevant. Consequently, a bidirectional LSTM was implemented on the features learnt by the CNN layer. Following the same time structure of the convolutional net, the bidirectional LSTM was designed to have 78 memory cells. The detailed LSTM architecture is shown in Fig. 5.

The outputs of LSTM cells from both directions are merged as the presentative features from each timestep. In that way, we preserve and refine the most discriminative features for the final emotion expression in terms of spatial as well as temporal information, which is the fundamental step for the following attention interface.
ATTENTION MODEL

Attention models have represented significant progress in many applications, such as machine translation [20], image captioning [21] and description [22]. Attention mechanisms allow neural networks to focus on the most pertinent piece of information, and predict accordingly attributes to this success. To be trained with the neural network, attention algorithms should be differentiable. A direct approach when classifying sequences is to give different weights to every sequence output based on its expected relevance. Fig. 6 illustrates the process of the attention interface we used for emotion recognition.

Following the same approach presented in 13, our attention mechanism gives different weights to each element in the sequence of outputs from the bidirectional LSTM. The attention algorithm is based on the scores obtained through the dot-product of the LSTM outputs and an attention vector (\( w_i \)). The attention vector is trained together with the neural network through the backpropagation algorithm. The score is subsequently fed into a SoftMax to produce the attention distribution (i.e., weights) (\( a_i \)) along all the output sequence. The distribution of these attention weights suggests to what extent different segments in the input utterance contributes to the target emotion recognition.

PERFORMANCE EVALUATION

Data Preprocessing

Attention model should be able to take care of the silence part of utterances, but considering the complexity of IEMOCAP, removing silence part [23] was still considered necessary so that attention model can pay more attention to other more specific parts of the utterance. In this paper, we use WebRTC Voice Activity Detector (VAD) provided by Google, for which a very easy-implemented interface has already been well-developed in Python [24]. An angry utterance sample after silence removal is shown in Fig. 7. We can observe that all silence part can be almost perfectly removed without destroying any original information structure.

![Figure 6. The architecture of attention model for emotion classification.](image-url)
Figure 7. An angry utterance with VAD.

Table 1. Experimental Setup.

<table>
<thead>
<tr>
<th>Network</th>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional layer</td>
<td>Number of CONV layers</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Convolutional filter size</td>
<td>3*3</td>
</tr>
<tr>
<td></td>
<td>Number of filters</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Max Pooling filter size</td>
<td>2*2</td>
</tr>
<tr>
<td></td>
<td>Activation function</td>
<td>Leaky RELU</td>
</tr>
<tr>
<td>Bidirectional LSTM layer</td>
<td>Number of cells</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>Activation function</td>
<td>tanh</td>
</tr>
<tr>
<td>Attention Layer</td>
<td>Dimension of attention distribution</td>
<td>256</td>
</tr>
<tr>
<td>Optimization (SGD)</td>
<td>Learning rate</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Decay</td>
<td>1e-6</td>
</tr>
<tr>
<td></td>
<td>momentum</td>
<td>0.8</td>
</tr>
<tr>
<td>Model training</td>
<td>Number of epoch</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Batch size</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 2. Weights of cost function for four emotions.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>0.95</td>
</tr>
<tr>
<td>happy</td>
<td>1.92</td>
</tr>
<tr>
<td>neutral</td>
<td>0.68</td>
</tr>
<tr>
<td>sad</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Table 3. Accuracy comparison among different architectures.

<table>
<thead>
<tr>
<th>Model</th>
<th>WA</th>
<th>UA</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN without Attention model</td>
<td>60.27%</td>
<td>49.96%</td>
</tr>
<tr>
<td>RNN with Attention model</td>
<td>61.13%</td>
<td>49.66%</td>
</tr>
<tr>
<td>CRNN without Attention model</td>
<td>62.36%</td>
<td>50.66%</td>
</tr>
<tr>
<td>CRNN with Attention model</td>
<td>64.08%</td>
<td>56.41%</td>
</tr>
</tbody>
</table>
In addition, we applied \(z\)-normalization to all utterance’s spectrograms with respect to the mean and standard deviation of spectrograms of neutral emotion in the training dataset. It can eliminate the variation of the characteristics of speakers to some extent and improve the final accuracy.

**NETWORK SETTINGS**

The parameter set-up for the whole experiments are shown in the following table. To balance the number of classes, we weighted the cost function based on the number of utterance for each emotion in the training dataset. Table 2 shows the weights when Session 2 is used as the validation dataset. We can observe that happy, the emotion with the larger number of training examples, has the highest weight.

**EXPERIMENTAL RESULTS**

We first tried to evaluate the efficiency of using CNNs for extracting discriminative and hierarchical features. Therefore, we compared our CNN + RNN model with a simple one HSF + RNN in which the first CNN layer was replaced by simple high-level statistical functions (HSF). The mean and standard deviation of spectral features within 32 adjacent short-term frames were used as HSF.

Therefore, in the HSF + RNN model the input dimension to RNN was 78*256. In both cases the configuration for the bidirectional LSTM was the same.

To assess the efficiency of the attention mechanism, we also compared the model using the attention weights with another one using simple temporal mean pooling; that is using the average of the RNN output sequence.

The results in terms of WA (weighted accuracy), which means the overall accuracy in validation dataset, and UA (unweighted accuracy), which is the average accuracy among four emotions, are shown in Table 3. In all the cases IEMOCAP session 2 was the validation dataset, so they correspond to speaker independent tests.

**SUMMARY**

In this paper, a novel approach based on the combination of distributed Convolutional and Recurrent Neural Networks (CRNN) together with an attention mechanism has been proposed. Experimental results give 64.08% weighted accuracy and 56.41% unweighted accuracy for IEMOCAP dataset, which, to the best of our knowledge, outperforms any other previous result on this dataset. The experimental results also indicate that CNN can extract more discriminative features from the raw spectrogram than the general statistical properties. We have also verified that attention model can improve the performance of SER systems.

In the future, we plan to continue our emotion recognition research but in a multimodal manner, and switch current four-emotion classification task into regression prediction on arousal, valence and liking. We believe multimodal emotion recognition can provide us a more comprehensive and accurate emotion detection performance.
ACKNOWLEDGEMENTS

This research was financially supported by Estudios y Soluciones Avanzadas de Ingeniería (ESAI).
Corresponding author: Yawei Mu, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, P. R. China.

REFERENCES


