SPEECH EMOTION RECOGNITION BASED ON DEEP RESIDUAL CONVOLUTIONAL NEURAL NETWORK

Chen Jin, A.I. Sherstneva, I.A. Botygin
Tomsk Polytechnic University, Russia, Tomsk

Abstract: Aiming at the accuracy of speech emotion recognition today, a one-dimensional deep convolutional neural network model with residual network is proposed. We extract Mel-frequency cepstral coefficients from sound files as input, and introduce a residual mechanism into the neural network to improve network training speed and recognition accuracy. The recognition accuracy of the model on Savee dataset and RAVDESS dataset is 98.26% and 97.30%.

Key words: Speech emothon recognition; Deep learning; Convolution neural network; Resnet

1. Introduce

Speech is one of the most important ways people communicate. A speech signal is a complex signal that contains information about the message, speaker, language, emotion, etc. In dialogue, non-verbal communication carries important information, which combined with different emotions, may lead to different semantic understandings of the same textual information [1]. Compared with other traditional classification methods for selecting features, deep learning models automatically extract the same features as their complex models [2]. For example, it can be applied in emotion recognition with important properties inherent to a particular sound file [3].

One of the most popular deep neural networks today is the Convolutional Neural Network (CNN). It consists of convolutional layers, nonlinear layers, pooling layers and fully connected layers [4]. The general deep convolutional neural network is constantly increasing its number of layers, but with the discovery of research, deeper neural networks are difficult to train, and may lead to the decline of training performance and the increase of training loss [5]. In order to prevent this shortcoming from causing problems with the network model, deep residual learning, called Residual Neural Network, or ResNet for short, was proposed in [6].

In this paper, we introduce residual neural network and propose a deep residual neural network model based on convolutional neural network. At the same time, we use the reciprocal Mel frequency extracted from the sound file as input, and use it to test the RAVDESS and SAVESS databases through the training of the model [7, 8].

2. The proposed model

In the constructed framework, we use a deep neural network model to extract MFCC spectral features from speech signals and perform classification. This model is based on the Convolutional Neural Network model, which combines Dropout, Batch Normalization and activation of 1D convolutional layers. Then, multi-layer residual blocks are added to construct a deep residual neural network model. The model is shown in Table 1.

Table 1. Network Model

Number	Name	Parameter
1	Conv1D	filters=256,kernel_size=8 padding=same
2	Activation	relu

3	Conv1D	filters=256,kernel_size=8 padding=same
4	BatchNormalization	----
5	Activation	relu
6	Dropout	0.2
7	MaxPooling1D	pool_size=8
8	(Conv1D+Activation) *3	filters=128,kernel_size=8 padding=same ‘relu’
9	Conv1D	filters=128,kernel_size=8 padding=same
10	BatchNormalization	----
11	Activation	relu
12	Dropout	0.2
13	MaxPooling1D	pool_size=8
14	(Conv1D+Activation) *2	filters=64,kernel_size=8 padding=same ‘relu’
15	Residual Block*10	channels_in =64,kernel_size=8
16	Dense *9	

2.1 Residual block

Figure 1 shows the basic architecture of the residual block in our study. This block contains one linear layer and two relu layers, and two or three convolutional layers. The number of convolutional layers depends on the number of feature layers. Consistency of the number of conv1 and conv2. If the feature layers are different, the input data needs to go through a convolutional layer to convert the consistency of the output size before the addition operation can be performed. If the feature layers of the two convolutional layers are the same, the input is assumed to be x, otherwise the output through conv1 is x, and the output feature layer of conv1 is the same as conv2. The first input x goes through the first convolutional layer, then the relu layer, the third conv2 convolutional layer, and finally a relu layer. The final output is added to x, and the result of this addition is the output of the entire residual block.

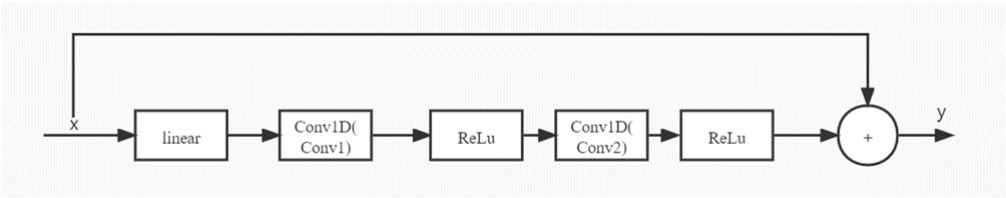

Figure 1. Residual Block

2.2 Convolutional layer

The convolutional layer is the core part of the convolutional neural network, and each layer is composed of several convolutional units (convolution kernels), which are used to perform convolution operations on the input data and generate corresponding feature maps.

Assuming that the input data is x(i,j), each convolution kernel is defined as w(i,j) of size a*b, and then the two implement the convolution operation, defining the convolution output as z(i,j) . Then there are:

$$z(i,j) = x(i,j) \times w(i,j) = \sum_{s=0}^{a-1} \sum_{t=0}^{b-1} x(s,t) \cdot w(i-s, j-t) \quad (1)$$

2.3 ReLU layer

The linear rectification layer, or ReLU for short, uses the linear rectification function:

$$f(x) = \max(0, x) \quad (2)$$

as an activation function. It can enhance the nonlinear characteristics of the entire neural network of the decision function kernel, and it does not change the convolution layer itself, making the entire model better generalization.

2.4 Maxpooling layer

Pooling is another important concept in convolutional neural networks, which is actually a form of downsampling. Among them, max pooling is the most common. It divides the input image into several rectangular regions and outputs the maximum value for each subregion. The pooling layer will continuously reduce the size of the data space, so the number of parameters and the amount of computation will also decrease, controlling overfitting to a certain extent.

3. Experiments and Results

In this experiment, we extract features from sound files, which are then used as input to train a deep residual network model to obtain a predictive model capable of emotional speech recognition. The process is shown in Figure 2:

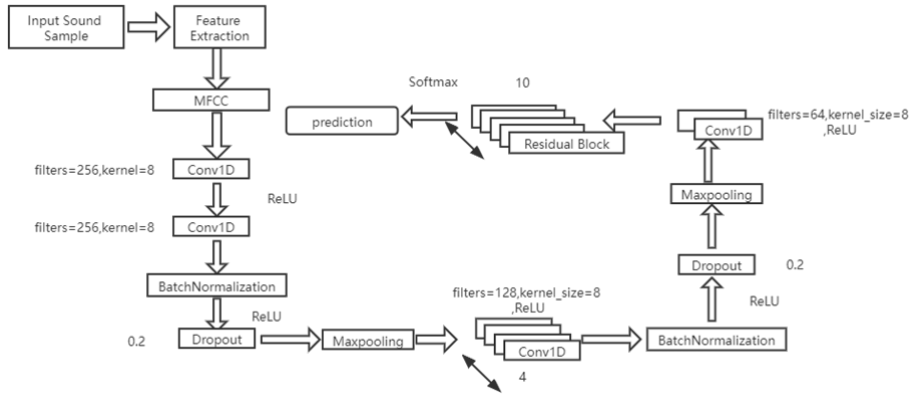

Figure 2. Experimental process

3.1 Dataset

We used two different audio databases, RAVEDESS [8] and SAVEE. Both are widely used by researchers for emotion recognition.

The RAVDESS dataset contains audio and video recordings of 12 male and 12 female actors pronouncing English sentences with eight different emotional expressions, including sad, happy, angry, calm, fearful, surprised, neutral, and disgusted.

The SAVEE dataset records four native English-speaking men who pronounce English sentences with different emotions, including anger, disgust, fear, happiness, sadness, surprise, and neutral. The textual material consists of 15 TIMIT sentences per emotion: 3 common sentences, 2 emotion-specific sentences, and 10 generic sentences, each different and phonetically balanced. 3 common and 2*6 = 12 emotion-specific sentences were recorded as neutral, giving 30 neutral sentences.

3.2 Feature extraction

Feature extraction plays a crucial role in the successful training of machine learning models. We choose Mel Frequency Cepstral Coefficients (MFCCs) as the input of deep learning models.

MFCC is widely used in the field of sound classification and speech emotion recognition [9, 10]. It imitates the inherent sound frequency reception pattern of human to a certain extent, and is more suitable as the input feature of speech emotion recognition model.

The feature extraction steps of MFCC are as follows:

(1)Framing the speech signal

(2)Power Spectrum Estimation Using the Periodogram Method

(3)Filter the power spectrum with a Mel filter bank, calculating the energy in each filter

(4)log the energy of each filter

(5)Perform DCT transform

(6)Keep the 2-13th coefficients of the DCT and remove the others

3.3 Experimental parameters

Our model parameters are shown in Table 1, the model uses RMSProp optimizer, and sets the learning rate to 0.0001 and the attenuation rate to 1e-6. Finally, the experiment is carried out. At the same time, the loss function uses categorical_crossentropy.

3.4 Result

Use this model to conduct experiments on two models, RAVDESS and SAVEE, respectively, and obtain their accuracy as shown in Table 2.

Table 2. Model accuracy on different datasets

Dataset	Accuracy
RAVDESS	97.30%
SAVEE	98.26%

At the same time, the changes of the model with the training accuracy are shown in Figures 3.

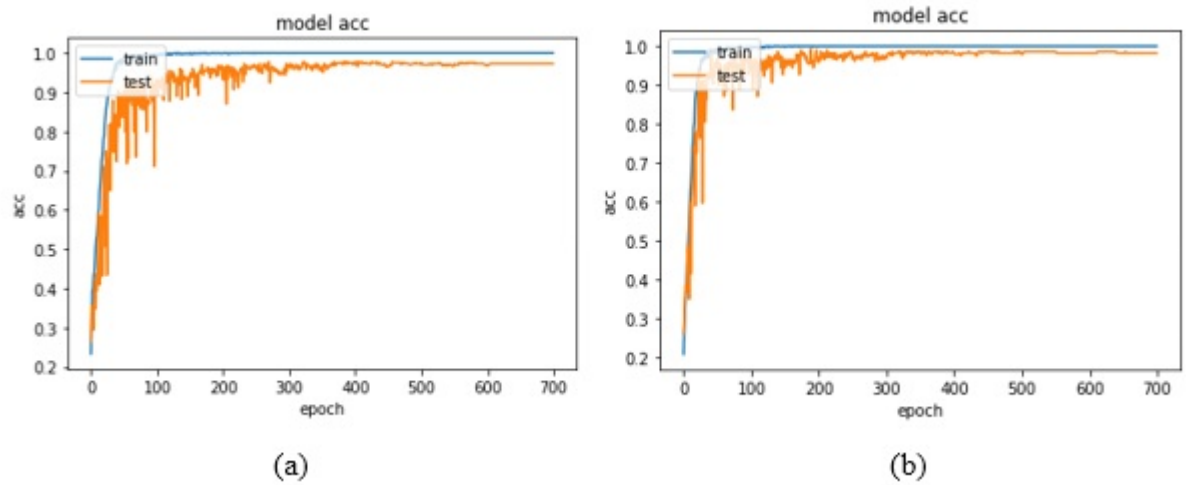

Figure 3. (a) — SAVEE, (b) — RAVDESS

4. Conclusion and Discussions

This work aims to introduce a residual neural network on the basis of deep convolutional neural network, and construct a one-dimensional deep residual neural network architecture, which enables it to recognize speech emotion signals in speech signals. After the residual neural network is introduced,

a deeper neural network can be constructed, avoiding the disadvantage of increasing the loss value of the training result due to the deepening of the network structure, and at the same time improving the recognition accuracy of the model.

For this deep convolutional neural network model, although the introduction of residual neural network has deepened the depth of its network model, there are still doubts whether it is possible to combine ResNet and GoogleNet to build a multi-scale depth convolution with coexistence of width and depth model, and applied to the recognition of speech emotion signals.

REFERENCES

1. Koolagudi, S.G., Rao, K.S. Emotion recognition from speech: a review. *Int J Speech Technol* **15,** 99–117 (2012). https://doi.org/10.1007/s10772-011-9125-1

2. Anagnostopoulos, CN., Iliou, T. & Giannoukos, I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artif Intell Rev* **4 3 ,** 155–177 (2015). https://doi.org/10.1007/s10462-012-9368-5

3. W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699.

4. S. Albawi, T.A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556.

6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385.

7. Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun, "An efficient MFCC extraction method in speech recognition," *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2006, pp. 4 pp.-, doi: 10.1109/ISCAS.2006.1692543.

8. 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.

9. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

10. Stevens, S. S. (1936). A scale for the measurement of a psychological magnitude: loudness. *Psychological Review, 43*(5), 405–416. https://doi.org/10.1037/h0058773