Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Research on Speech Emotion Recognition Analysis Based on Deep Learning

Ailiang Yin*, Chunhao Li

School of Electronic Information, Southwest Minzu University, Chengdu, China
* **Corresponding author**: Ailiang Yin

Abstract: This paper combines two aspects of feature selection and building deep neural networks to carry out targeted research to improve recognition accuracy. Firstly, speech preprocessing techniques are introduced to extract the speech spectrogram and lay the foundation for building the speech emotion recognition network model study. The focus is on building a speech emotion recognition network model based on residual network improvement and comparing experiments with AlexNet model network and ResNet-18 network model.

Keywords: Speech spectrogram; Residual network; Speech emotion recognition.

1. Introduction

In addition to the basic semantic information, speech signals also contain a variety of complex human emotions, which play a very important role in human communication and life. The importance of speech emotion recognition in the field of artificial intelligence is self-evident. With the development of science and technology, artificial intelligence is becoming more and more mature, but the development of self-awareness and self-awareness of artificial intelligence is slow, which hinders the natural interaction process between human and machine. One important way to solve the self-awareness and self-awareness of AI is to make the machine understand human's emotional state, so that the intelligent machine has the ability of emotion.[1][2] This is an important way to address the self-awareness and self-awareness of AI is to enable machines to understand human emotional states, so that intelligent machines have emotional capabilities.

In recent years, deep learning algorithms have achieved excellent results in the field of natural language, and the classic result is the recurrent neural network (RNN).[3] RNNs have been shown to be effective in solving serialization problems, making use of contextual information, but they suffer from gradient explosion and vanishing problems in the solution process, and are ineffective in processing long texts. However, RNNs suffer from the gradient explosion and disappearance problems during the solution process, and are not effective for long text. The problem has been effectively solved by the proposed Long-Short Term Memory Neural Network (LSTM), which is suitable for solving long sequential problems by adding three "gates" in the hidden layer to control cell states.[4][5] It is suitable for solving long sequence problems.

Speech emotion recognition, as a crucial technology to realize human-computer interaction, has a wide range of applications in many fields with natural human-computer interaction needs. For example, we can track the emotional changes of depression patients to understand the specific condition of patients from these emotional changes, so as to make corresponding diagnosis and treatment; we can monitor the emotional changes of each student in real time in the teaching classroom to find out the interest of each student in the lecture content from these emotional changes, so as to make timely adjustments to the lecture content and lecture style.[6] The class is monitored in real time to detect each student's interest in the lecture content, so that adjustments can be made to the lecture content and lecture style.

2. Related work

This paper adopts the features of the speech spectrum map as the input data of the network structure, which solves the problem that the data structure is not similar to the two-dimensional structure of image data. The traditional convolutional neural network has serious data loss problems in information transfer, and the gradient disappears or explodes as the number of layers of the network deepens.ResNet network can effectively solve the above problems, and this paper will use ResNet network as the backbone network structure.

2.1. Speech pre-processing technology

Emotion feature extraction is important in speech emotion recognition. Currently, there are many effective emotion features commonly used in the field of speech emotion feature research, and they express the emotions embedded in speech at different levels respectively. In this thesis, the extracted speech spectrograms are used as the dataset for model training. The speech spectrogram represents the speech signal in visual form and contains many essential information of the speech signal, such as: fundamental period, amplitude, etc. In a speech spectrogram, a two-dimensional plane is used to represent three-dimensional information, the horizontal axis shows the time and the vertical axis shows the frequency, and the intensity of a frequency at a certain moment can be indicated by the color shade or the gray scale of the point, and the energy value can also be indicated by the color shade, the darker the color of a point means the stronger the energy at that place.[7][8] The darker the color of a point, the stronger the energy at that point. The specific speech spectrum extraction technique is shown in Figure 1.

The speech spectrogram is obtained by Fourier analysis of the speech signal. First, the discrete speech signal $x(n)$ is represented as $x(m)$, n = 0, 1 ••• , N-1 after framing, m is the sample point number within a frame, n and N denote the frame number and frame length, respectively. The short-time

Fourier transform of the signal $x(n)$ is shown in the following equation.

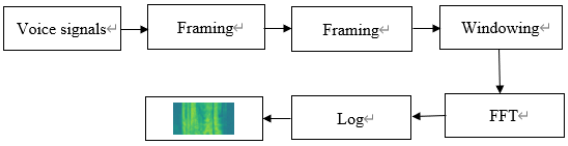

Figure 1. The process of extracting the language spectrum map

$$x_n\left(e^{jw}\right)=\sum_{m=-\infty}^{\infty}x(m)\cdot w(n-m)\cdot e^{-jwm} \quad (1)$$

w(n) is the window function, and the discrete time domain Fourier transform and discrete Fourier transform of x(n) are as follows.

$$\begin{aligned} X\left(n,e^{jw}\right) &= \sum_{m=0}^{N-1}x_n(m)e^{-jwm} \\ X(n,k) &= \sum_{m=0}^{N-1}x_n(m)e^{-j\frac{2\pi km}{N}} \end{aligned} \quad (2)$$

where 0 ≤ k ≤ N - 1, X(n, k) denotes the short-time amplitude spectrum estimate of x(n), and the spectral energy density function P(n,k) at m can be expressed as:

$$P(n,k)=|X(n,k)|^2=(x(n,k))\times(conj(X(n,k))) \quad (3)$$

The horizontal coordinate is n and the vertical coordinate is k. The values of P (n, k) are represented in color or gray, and finally the speech spectrum diagram is obtained. The speech spectrogram is generally a gray or colored two-dimensional image, and each point contains the corresponding time, frequency, energy and other information of the speech signal. the speech spectrogram of a speech segment of anger class in CASIA Chinese emotion dataset is shown in Figure 2.

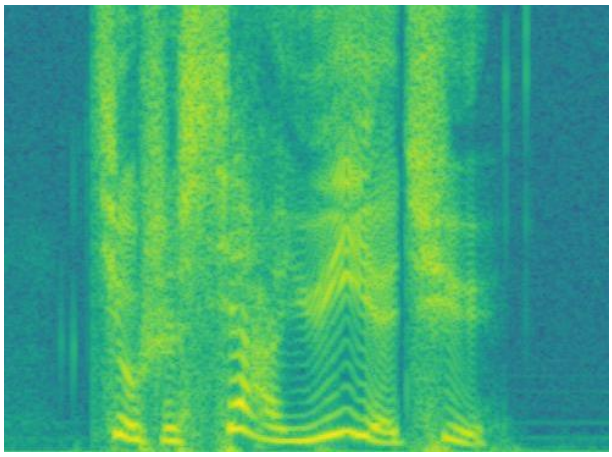

Figure 2. The emotional spectrum of the anger category in CASIA Chinese emotional database

From Figure 2, we can clearly see many horizontal stripes, i.e., sound patterns, which are formed by the aggregation of dark-colored pixels, i.e., by the extension of the duration points of time, indicating that the frequency component at that moment has a large proportion and is easily perceived. The greater the proportion of stripes represents the greater the proportion of effective information in the whole speech signal, and the more obvious the influence on the speech emotion recognition rate. There is a lot of information that can be obtained from the speech spectrogram, such as the change of sound intensity in a certain period of time, the frequency distribution and sound intensity strength of the whole speech, etc. For different emotions of speech, there is a certain degree of difference in the speech spectrogram, and through these differences, different emotions can be classified[9][10] .

2.2. Att-Res speech emotion recognition network model construction based on residual network improvement

In this paper, the language spectrum map is selected as the primary feature for network input. Compared with the combination of artificial features, the language spectrogram features bring more comprehensive and rich information, which can be further extracted by deep neural network for deep features. At the same time, the spectrogram features bring redundant information with emotionally irrelevant information due to the comprehensive speech information they contain, which brings impact on the final recognition accuracy. The attention mechanism simulates human brain vision, learns the weight distribution from the features and imposes it on top of the original features, focuses on the effective features and weakens the invalid features, which can effectively improve the feature extraction efficiency.

The structure of the spatial visual attention mechanism is shown in Figure 3. The spatial attention module first performs the input feature map by channel$F$ Maximum pooling and average pooling operations are performed. Maximum pooling is to keep the maximum value for the pooling window. The average pooling is to keep the overall data features for the pooling window and calculate the average value. The two pooling results are concatenated by channel to represent the input holistically, and then a 1×1 convolution layer for feature extraction by convolution, followed by using the $softmax$ activation function to convert the obtained single-layer feature maps into probability distributions to obtain the attention mechanism scores for each element of the input features. The larger the attention score, the more the information of the corresponding element is worthy of attention in identifying the sentiment category, and conversely, the smaller the attention score, the less impactful and less valuable the information of the corresponding element is in identifying the sentiment category, and can be ignored. Finally, the attention score matrix and the input values$F$ A dot product operation is performed to adjust the model to the input$F$ importance of different spatial locations.

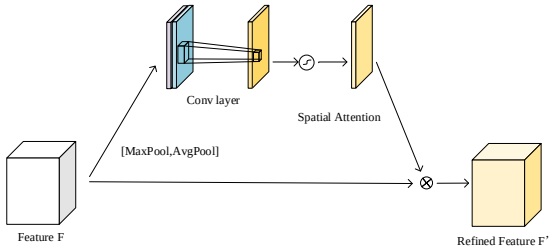

Figure 3. Spatial attention mechanism

In this paper, we combine the residual network and introduce the spatial attention mechanism in the residual block to build the Res-A module, aiming at the training process when the network pays more attention to the sentiment features and ignores the redundant information. the structure of the Res-A module is shown in Figure 4.

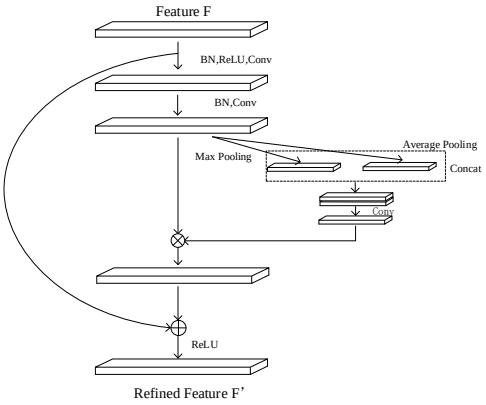

Figure 4. Res-A module

In this paper, based on the Res-A module, we design and build the att-Res network structure with 15 layers, as shown in Figure 5. In the neural network model, it is experimentally proven that a smaller convolutional kernel can be used instead of a larger convolutional kernel to obtain better training results, and Inception network structure proposes to use 1× 1 and 3× 3 convolutional kernels can be used instead of 5× In the Inception network structure, it is proposed that 1 1 and 3 3 convolutional kernels can replace 5 convolutional kernels and obtain better results. In the development of neural networks, 3 × 3 convolutional kernel size is the most commonly used convolutional kernel in the development of neural networks, and it appears many times in classical networks such as Transformer, VGGNet, Inception, etc. The convolutional kernel size of 3 is also chosen for att-Res network.× For network models such as Resnet for image processing, the data input to the network appears to pass through 64 or more convolutional kernels, and the convolutional kernel size is larger, such as 7× 7 or 5× 5 convolutional kernels to speed up the training efficiency. The size of the speech spectrogram data varies with the length of the speech, so in order to standardize the size, the size is modified to 128 after extracting the spectrogram for each speech signal using the cv2.resize function.× 128 and then input to the network for training. The spectrogram size is small compared with the image size, so the input layer of the att-Res network is chosen to have 32 3× 3 convolutional kernels. The speech sequences are temporal in nature, and the speech spectrogram is also stitched according to the time axis of the speech frame data. Therefore, the convolution step in the input layer of the att-Res network is set to 1 to extract the continuous shallow links. Subsequent data pass through a normalization layer, a maximum pooling layer, a constant link block, and again through data passing through two Block-A blocks, each Block-A consisting of a convolutional layer, a normalization layer, a maximum pooling layer, and a Res-A block. One of the convolutional layers contains 64 or 128 3× The final features are identified by a fully connected layer and a softmax classifier for sentiment classification. The network layer activation functions all use the ReLU function.

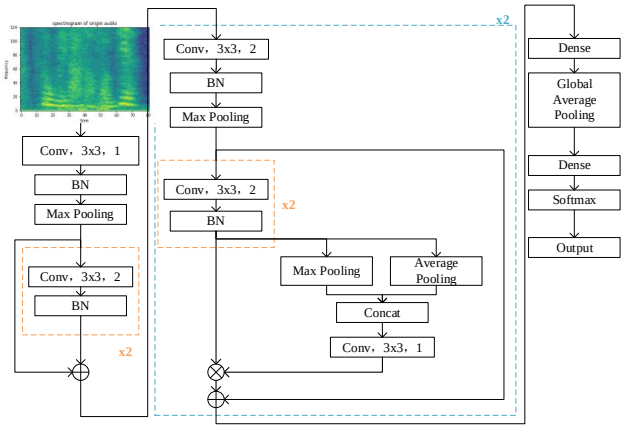

Figure 5. Att-Res network structure

3. Experimental setup, results and analysis

This paper selects CASIA Chinese sentiment dataset recorded by the Institute of Automation of Chinese Academy of Sciences, and conducts training and testing on this dataset. The training framework adopts Nvidia GTX2060 graphics card, deep learning Pytorch environment, which supports Python, deep learning model training, its package of libraries is complete and complete, easy to quickly build models, support GPU acceleration This environment supports Python, deep learning model training, is complete, easy to build models quickly, supports GPU acceleration, and enables fast scientific computing, which provides the conditions for smooth experiments.

3.1. Experimental setup

In this paper, the CASIA dataset is used for the experiments. The experiments are mainly set up as a set of ablation experiments to verify the effectiveness of the proposed module and a set of comparison experiments to verify the effectiveness of the proposed network model. Block-A is composed of one layer of convolution, one layer of normalization, one layer of maximum pooling and one module of Res-A. If Block-A is in the input layer of the network, the step size of the convolution layer is 1, otherwise the step size is set to 2 to speed up the training.

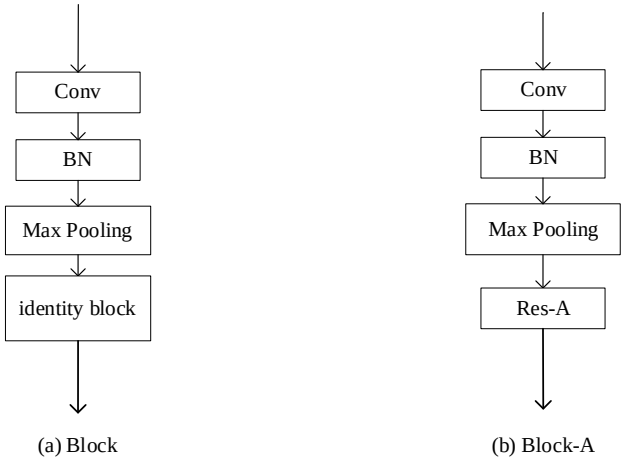

Figure 6. Block and Block-A

In order to better analyze the impact of Res-A module on the performance of the overall network model, this paper uses the att-Res0 network constructed by three Blocks as the

backbone network to extract features, and gradually replaces Blocks with Block-A to constitute four network structures, att-Res1, att-Res2, att-Res and att-Res3, and the five network structures are shown in Figure 7 shows.

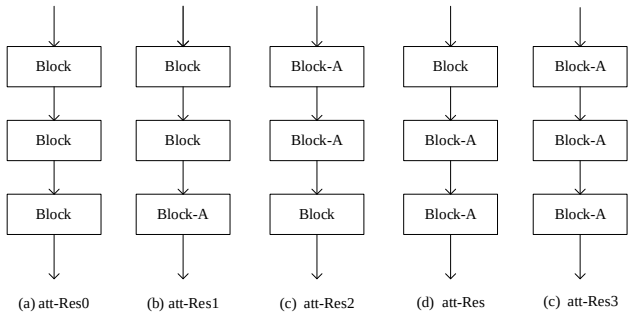

Figure 7. att-Res0, att-Res1, att-Res2, att-Res and att-Res3 network structure

To discuss the effects of the two approaches, Res-A module and adding spatial attention mechanism directly after feature extraction, on the network model, this paper constructs an att-Res0-att network based on the att-Res0 network, aiming to add the spatial attention mechanism before the global averaging layer to form a comparison experiment with the att-Res1 network model. att-Res0-att network is shown in Figure 3- 3 shows. In this paper, five network structures are ablated with the baseline network to analyze the effect of Block-A modules on the accuracy of emotion recognition and the effect of the number and position of Block-A modules on the accuracy of emotion recognition.

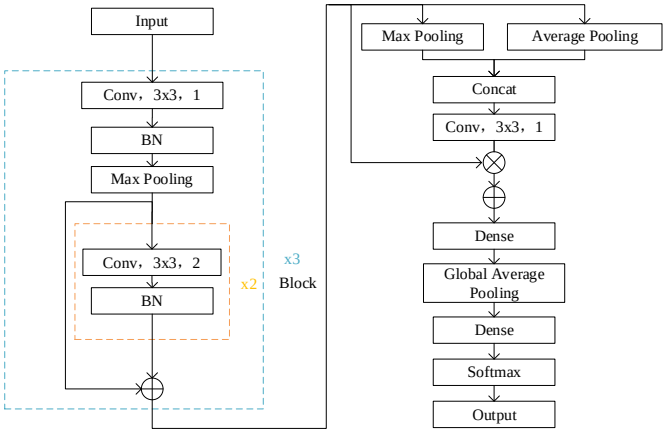

Figure 8. Att-Res0-att network structure

In order to better analyze the effectiveness of the network model on the speech emotion recognition task, this paper further analyzes the effectiveness of att-Res network for small sample English speech emotion recognition by comparing the experimental results with ResNet-18, SVM, AlexNet and other network models.

In the experiments, the data set is randomly divided according to the ratio of 80% being the training set and 20% being the test set. The Adam optimizer was chosen to continuously optimize the learning network parameters, and a total of 25 iterations were trained.

3.2. Results and Analysis

The test set accuracies for att-Res0, att-Res0+att, att-Res1, att-Res2, att-Res and att-Res3 are shown in Table 1.

Comparing the experimental results in Table 1, we can find that the improvement effect of the att-Res0+att model on recognition accuracy is significantly lower than that of the att-Res1 network model, indicating that the Res-A layer is better than the direct addition of the attention mechanism layer in extracting effective features, which verifies the effectiveness of the Res-A layer. The recognition rates of the network models with the Block-A module added are all better than those of att-Res0 composed of the Block module only. The biggest difference between the Block-A module and the Block module is that the Res-A layer proposed in this paper replaces the constant residual layer in the residual network, which effectively improves the network sentiment recognition accuracy by entering spatial attention. In the four network models composed of a mixture of Block-A modules and Block modules, the recognition accuracy increases as the number of Block-A modules increases. However, the recognition accuracy decreases when the number of Block-A modules is 3. The CASIA data are small sample data, and the excessive introduction of attention mechanism tends to make the network in an overfitting state, which reduces the model effect. In the network consisting of two layers of Block-A modules and one layer of Block modules, we explored the effects of introducing Res-A layers at different locations. This is because the number of deeper features extracted when the network is deeper has more semantic information, and then adding the attention mechanism can further extract the emotional features.

Table .1 Ablation experiments

Experimental group	Dataset	Methods	Accuracy
1	CASIA	att-Res0	57.2%
2		att-Res0+att	58.7%
3		att-Res1	64.4%
4		att-Res2	67.3%
5		att-Res	69.8%
6		att-Res3	66.1%

Finally, to verify the effectiveness of the att-Res model extracted in this paper for speech emotion recognition, this subsection compares five commonly used network models for the CASIA dataset, where experimental groups 4, 5, and 6 all use the speech spectrogram as the input to the network. The experimental results are shown in Table 2.

Table 2. Speech emotion recognition on CASIA dataset

Experimental group	Dataset	Methods	Accuracy
1	CASIA	CNN	46.9%
2		LSTM+CNN	49.2%
3		SVM	50.2%
4		AlexNet	51.7%
5		ResNet-18	50.3%
6		att-Res	**69.8%**

Compared with the ResNet-18 network structure, the recognition accuracy of the improved att-Res network based on the residual network is higher because the incorporated spatial attention mechanism can autonomously learn the distribution of emotion features and weight them by back-propagation, thus achieving the effect of reducing the interference of redundant information on the recognition accuracy. Finally, comparing all the experimental results in Table 3-2, the improved att-Res network of the model in this paper can improve the accuracy of the network model recognition in the comparison experiments. att-Res network

model improves the speech emotion recognition rate by 22.9% compared with the CNN network model and 19.5% compared with the ResNet-18 network, which verifies the model's effectiveness.

4. Summary

In this paper, we introduce attention mechanism into constant residual block, propose a Res-A module based on constant residual block, design six network models of att-Res0, att-Res0+att, att-Res1, att-Res2, att-Res and att-Res3 for this module, verify the effectiveness of this module with ablation experiments, and the experimental results prove that the Res -A module can effectively improve the speech emotion recognition rate.

Finally, this paper proposes an att-Res speech emotion recognition network structure based on residual block improvement based on the Res-A module, and trains the model on CASIA dataset. By comparing with four advanced network models and a traditional classifier on the same dataset, the results demonstrate that the att-Res network model improves speech emotion recognition to a certain extent and verifies the feasibility of the model.

Acknowledgments

This work was supported by the Southwest Minzu University for Nationalities Graduate Student Innovative Project No. (YB2022753) fund.

References

[1] Hong Zhaogin,Wei Chenyang,Zhuang Yuan,Wang Ying,Wang Yiting,Zhao Li. Speech emotion recognition and personality analysis based on deep neural network[J]. Information Technology Research,2020,46(01):48-53.

[2] Tao, H.W., Cha, C., Liang, R.Y., et al. A speech spectrogram feature extraction algorithm for speech emotion recognition. Journal of Southeast University (Natural Science Edition),2015,45(5):817-827.

[3] Irsoy O, Cardie C. Deep recursive neural networks for compositionality in language[J]. Advances in neural information processing systems, 2014(3).2096-2104.

[4] Shah A, Bhowmik T. A Comparative Study on MFCC and Fundamental Frequency Based Speech Emotion Classification[C]//International Conference on Distributed Computing and Internet Technology. springer, Cham, 2022: 173-184.

[5] Han WJ, Li HF, Ruan HB, et al. A review of research advances in speech emotion recognition[J]. Journal of Software,2019,25(1):37-50.

[6] Qiu Xue. Research on multimodal fusion for student emotion recognition and state analysis[D]. Supervisor: Li Mingyong. Chongqing Normal University,2021.

[7] Gao L, Qi L, Chen E, et al. Discriminative multiple canonical correlation analysis for information fusion[J]. IEEE Transactions on Image Processing, 2017, 27(4): 1951-1965.

[8] New T L, Foo S w, De Silva LC. Speech emotion recognition using hidden Markov models[J]. Speech communication, 2003, 41(4): 603-623.

[9] Liu, Jingjing, Wu, Xiaofeng. Multimodal sentiment recognition and spatial annotation based on long and short term memory networks [J]. Journal of Fudan (Natural Science Edition),2020,59(05):565-574.

[10] Chen Jia. Speech emotion recognition based on deep learning [D]. Nanjing University of Posts and Telecommunications, 2019. 000586.