Application of Improved Neural Network Strategy in Emotion Information Discrimination

Junlian Huang 1, Dongyan Zhao 1,a,*, Boxue Lv 2

1 Gengdan Institute of Beijing University of Technology, Beijing 101301, China
2 Vimicro Corporation, Beijing 100000, China

*Corresponding author e-mail: Huangjunlian@mail.gengdan.edu.cn

Abstract. Emotional information recognition is an important application in the field of artificial intelligence. This paper analyzes the existing speech emotion recognition system, based on the improved neural network strategy to improve the accuracy of speech emotion recognition, through the combination of convolution neural network (CNN) and long-term and short-term neural network (LSTM), the emotional data information is successfully recognized, and the better accuracy is obtained.

Keywords: Emotion Recognition; Neural Network; CNN; LSTM.

1. Introduction

With the rapid development of computer system, people pay more and more attention to efficient human-computer interaction. How to make the computer accurately perceive the user's emotion, provide the appropriate dialogue environment for the operator, reduce the obstacles between the operator and the machine as far as possible, and improve the human-computer efficiency to the greatest extent, has become an important development direction of the future computer [1].

At present, the technology route of human-computer interaction is changing from mechanical interaction to intelligent and humanized. The traditional human-computer interaction mode is keyboard, touch screen and so on. Although speech recognition, as a new human-computer interaction interface technology, has made some achievements in various fields of reality, its more rigid behavior mode can not meet the needs of modern people For the need of more intelligent and humanized interaction mode, the demand of speech emotion recognition becomes more urgent.

Voice is one of the most basic, direct and effective ways for human beings to transmit information to each other. Through voice, human beings can transmit information to the receiver in a direct and efficient way. Voice information not only contains the text information expressed in the actual pronunciation, but also contains the information to identify the identity of the speaker, as well as the emotional information of the speaker, such as happiness, anger, sadness and so on, Language is closely related to human intelligence, which makes speech communication have the highest intelligence level and the largest information capacity.

As early as the 1980s, the research on speech emotion recognition has begun. At that time, the main research direction was to use acoustic features to classify speech emotion. In 1990, Moriyama, a scientist, proposed a simple model based on the linear correlation between speech and emotion, and built a model...
that can collect user's speech and recognize user's emotion at the same time. At the same time, this system is also the first time for speech emotion recognition research to be commercialized [2].

In the late 20th century, R. Under the call and advocacy of Picard [3], researchers gradually realized that emotional intelligence is the crux of this series of problems, and the research field of "emotional computing" was established, which is a science to establish a harmonious man-machine environment by giving computers the ability to recognize, understand, express and adapt to people's emotions, and to make computers have higher and more comprehensive intelligence [4, 5, 6].

After entering the 21st century, the development speed of speech emotion recognition is accelerating with the research upsurge in the field of artificial intelligence. Many conferences and journals with this theme have been established and attracted attention. They started from the biennial meeting of affective computing and intelligent interaction in 2005 and the inter Speech Emotion in 2009. The annual challenge competition is famous in this field [7]. At the same time, more and more universities and scientific research institutions also pay attention to speech emotion recognition, and have achieved excellent results in this field. Speech emotion recognition has rich development space, its research results can be used in a variety of scenarios, such as in the treatment of timely detection of the psychological state of the treated, timely make targeted changes, so that the treatment can be carried out more smoothly; in the process of criminal investigation, judge the emotional state of the suspect to identify whether the object is lying; in addition, speech emotion recognition can also be used to use in electronic life products, such as mobile phones, cars, home appliances and many other fields, can greatly improve people's quality of life.

2. Convolutional Neural Network CNN

Fully connected neural network is a neural network structure that connects every neuron in N layer with all neurons in n-1 layer. Although it still has good recognition accuracy, it has very obvious disadvantages: firstly, fully connected neural network needs too many parameters. Secondly, the fully connected neural network is considered to be of equal importance for each point, but in the calculation and identification, there are a lot of elements that are completely unnecessary, which is bound to waste a lot of extra computing power. Finally, the depth of the fully connected neural network is. As we all know, the more layers, the stronger the expression ability of the neural network. Because the traditional SIGMOD function is used in the fully connected neural network, the problem of gradient vanishing is very easy to occur, which makes it difficult for the depth to exceed three layers, which further limits the ability of the neural network. Due to the disadvantages of fully connected neural network, researchers put forward the concept of constructing deep neural network with convolution neural network.

In order to solve the problem of too many unnecessary parameters in fully connected neural network, convolutional neural network puts forward the following solutions: first, the neurons in each layer do not connect with all the neurons in the upper layer, but only want to connect with a small part of the neurons in the upper layer, so that a considerable number of parameters are reduced and a very large amount of computation cost is reduced Quantity. Secondly, convolutional neural network further proposes the concept of weight sharing. The connection of a group of neurons can share a weight, instead of each neuron using different weights, which further reduces the number of parameters. At last, the concept of down sampling is put forward. After the calculation of each layer, the down sampling layer is used to reduce the number of samples in each layer, further reduce the number of parameters, and improve the robustness of the model is shown in Figure 1.

![Convolutional neural network](image)
3. **Long Short Time Neural Network LSTM**

Cyclic neural network has the defect of gradient disappearance, which makes it unable to capture long-distance influence. The long-term and short-term memory network LSTM invented by Hochreiter and Schmidhuber solves this problem [8].

Aiming at the defect that the traditional recurrent neural network is extremely sensitive to short-term input, LSTM adds the cell state (c) to maintain the long-term state on the basis of the recurrent neural network.

In addition, the gate structure is used to control the output of the signal. It includes: forgetting gate, input gate and output gate.

3.1. **Forgetting gate**

Calculation of forgetting gate:

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

Where: \( W_f \) is the weight matrix of the forgetting gate; \([h_{t-1}, x_t]\) links two vectors into a longer vector; \( b_f \) is the bias term of the forgetting gate, \( \sigma \) is the sigmoid function.

3.2. **Input Gate**

The second step is to decide what information can enter the cell. First, a sigmoid layer called input gate layer is needed to decide which information needs to be updated [9], and a tanh layer to generate a vector of alternative update content \( \tilde{C}_t \). After that, the information is combined to update the cell state.

Calculation of input gate:

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

Where: \( W_i \) is the weight matrix of the input gate; \( b_i \) is the offset term of the input gate.

\[ \tilde{C}_t = \tanh(W_c * [h_{t-1}, x_t] + b_c) \]  

\[ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \]  

3.3. **Output Gate**

The third step is to decide what value to output. The sigmoid layer will first decide which part of the cell state to output and send the cell state to the tanh layer for processing. After the tanh layer has finished processing, the cell state is multiplied by the output of the sigmoid gate, and finally the result is output [10].

Calculation of output gate:

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

\[ h_t = O_t * \tanh(C_t) \]

4. **CNN Speech Emotion Recognition Based on LSTM**

The data used in this paper is obtained from 10 sentences including neutral, anger, fear, happiness, sadness, disgust and boredom, which are simulated by five male and five female students. It contains 800 sentences, and the sampling rate is 16KHz, 16bit quantization. The corpus text follows the principle of semantic neutral and non-emotional tendency, requiring students to recall their real experience and brew emotions before performing a certain emotion, so as to enhance the authenticity of emotions. After 20 participants’ listening experiments, the recognition rate of 86.5% is achieved.

Because the original voice signal is continuous, and the data used by the computer is discrete, which brings great trouble to the process of data processing in the computer, so we need to transform the original analog data into discrete data in the form of sampling is shown in Figure 2.

![Figure 2. Original Voice Signal](image-url)
Because the sampling rate of the voice emotion database used in this neural network design is commonly used 16KHz, which provides considerable convenience for the sound preprocessing process of this design. The network used in this paper is based on long short memory network (LSTM) and convolutional neural network (CNN) to train and test speech emotion recognition.

Firstly, the input speech is extracted from the speech feature information, and the feature data is input into the long-term and short-term neural network for processing is shown in Figure 3.

![Figure 3. Structure of LSTM](image)

The structure of long-term and short-term memory neural network LSTM is a single layer neuron. LSTM combines the extracted context information into a feature matrix and outputs it to convolutional neural network CNN is shown in Figure 4.

![Figure 4. Structure of CNN](image)

CNN network consists of three conv processing modules and a fully connected network. Each iteration contains a convolution layer, a normalization layer and a pooling layer. The convolution layer is used to extract the matrix features, and the normalization layer outputs the matrix as a decimal between (0, 1) to reduce the amount of parameters, so as to reduce the amount of calculation. The pooling layer is responsible for reducing the length and width of the matrix, which is also used to reduce the amount of parameters. The matrix output by the processing module will be input to the full connection layer and output the results.

In this paper:
The convolution kernel used in conv1 has a width of 5 * 5 and generates 8 feature maps. The pooling layer does not reduce the length and width of the matrix. The convolution kernel used in conv2 is 5 * 5 in width, generating 16 feature maps. The convolution kernel used in conv3 is 3 * 3 in width, generating 32 feature maps. The pooling layer in conv2 and conv3 modules reduces the matrix to half. There are 64 neurons in the whole connective layer. After the whole connective layer processing, 7 results are output.

![Figure 5. Recognition Accuracy of Previous Algorithm](image)

As can be seen from Figure 5, the recognition effect of CNN speech emotion recognition strategy based on LSTM ranges from 62% to 83%, achieving good recognition effect.

5. Conclusion
Based on the neural network method, this paper proposes a strategy of combining convolutional neural network (CNN) with long-term memory network (LSTM). The context features of speech signal in time domain are extracted by long-term memory network. On this basis, the convolutional neural network is used to extract the emotional information features and complete the classification of emotional information. The simulation results show that the strategy is effective. The accuracy of speech emotion recognition is improved. Because the parameters of the neural network are very large, it means that the neural network needs a large number of data sets for training in order to avoid under fitting phenomenon, that is, the neural network learning for a certain feature is insufficient, which makes the neural network unable to correctly identify the feature, and eventually leads to the calculation error of the neural network for the whole input information. However, this paper still has a positive reference for the feature extraction and analysis of emotional data. The next step is to improve the recognition accuracy and expand the recognition ability.

References
[1] Li Chao. Speech emotion recognition based on semi supervised learning[D]. Tianjin Normal University, 2011.
[2] Chen Xiaodong. Speech emotion recognition based on convolutional neural network [D]. Guangzhou: South China University of Science and Engineering, 2015.
[3] Picard R W. Affective Computing[M]. Cambridge: The MIT Press, 1997.
[4] Schröder M. Emotional Speech Synthesis: A Review[C]. Proc. the 2001 Euro-pean Conference on Speech Communication and Technology. Aalborg, Denmark: Eurospeech, 2001: 561–564.
[5] Yankai Xu. Development of Ultrasonic Small Flow Measuring Instrument Based on STC89C52.
2019 2nd International Conference on Safety Produce Informatization, IEEE. Chongqing, Nov. 28-30, 2019, p:394-397.

[6] Han Wenjing. Research on key technology of speech emotion recognition [D]. Harbin: Harbin Institute of technology. 2013.

[7] YI Na; XU Jianjun; YAN Limei; HUANG Lin; Task optimization and scheduling of distributed cyber - physical system based on improved ant colony algorithm, Future Generation Computer Systems, 2020, 109: 134-148.

[8] Zhang Jian. Research on recurrent neural network language model technology in continuous speech recognition [D]. Zhengzhou: PLA University of information engineering, 2014.

[9] Yang Zhao, Jianjun Xu, Jingchun Wu. A New Method for Bad Data Identification of Oilfield Power System Based on Enhanced Gravitational Search-Fuzzy C-Means Algorithm. IEEE Transactions on Industrial Informatics, 15(11), NOV. 2019: 5963-5970.

[10] Xue Yuan. Research on Chinese word segmentation based on deep learning algorithm [J]. Computer products and circulation, 2019.

[11] Xu, J.J., Huang, L., Yin, S. et al. All-fiber self-mixing interferometer for displacement measurement based on the quadrature demodulation technique. Opt. Rev. 2018,25(1):40-45.

[12] Yang Fan, Yan Limei, Xu Jianjun, Li Hongyu. Analysis of optimal PMU configuration method based on incomplete observation. Concurrency and Computation-Practice & Experience. JUN,2019,31(12):e4835.

[13] Wang Chao. Research on convolutional neural network control strategy of manipulator based on deep reinforcement learning. Harbin: Harbin Institute of technology, 2018.