Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Research on Speech Emotion Recognition Method Based on ResSE_CNN1D

Yingcheng Zhang*

College of Southwest Jiaotong University, Chengdu, Sichuan province, China

* 3119127438@my.swjtu.edu.cn

Abstract. This article examines a speech emotion recognition (SER) technique based on the enhanced one-dimensional convolution neural network ResSE_CNN1D. With the artificial intelligence developing rapidly, SER has a profound impact in many areas. The model of this article is used to extract the characteristics of the input data through opensmile, and is sent into the ResSE_CNN1D model, which is eventually classified by the softmax activation function and obtains the final results. The key to this model is efficient learning of decimal sets and the rapid deployment in resource-constrained environments. The ResSE_CNN1D model improves the performance of the model by adding the residual connection and the SE module on the basis of cnn1d. This increasing the accuracy of the recognition and preventing the fitting problem. After the model was created, the study adopted the audio sampling and training of the casia data concentration. The final accuracy was 0.900, which increased the accuracy of 2.9 percent compared to the cnn1d method. And by the analysis of the relationship diagram of the confusion matrix and the accuracy and loss rate relative to the number of training, the model has a high robustness and effectively prevents the appearance of the fitting problem. And it also can achieve high precision and achieve a lightweight goal relative to less training.

Keywords: ResSE_CNN1D, opensmile, SER, CASIA.

1. Introduction

The need for human-computer interaction is growing as artificial intelligence technology develops at a rapid pace. Speech emotion recognition (SER) , as one of the most important technologies in human-computer interaction, can be identified by extracting the characteristics of the input speech to build the mapping relationship between characteristics and emotions, and finally realize the sense of emotion in the input speech. The technology has broad application prospects in medical, education, business and so on[1]. In medicine, the patient's voice can be judged by the mood of the patient, so that the psychiatrist can better treat the disease. In education, when you are in the Internet class, you can judge the state of their class by the voice of the teacher and the student. In business, intelligent customer service can adjust different solutions through customer voice emotional state and satisfy customers.

For the first time, Kim employed CNN to categorize text emotions and discovered that CNN outperformed other conventional machine learning systems[2]. Zhao xiaolei et al. feed the support vector machine with depth learning and artificial statistics features to get a good recognition rate[3]. However, most of these models are focused on the ascension of universality, this model is dedicated to the efficient learning of decimal sets and the rapid deployment in resource-constrained environments. Besides, this paper puts forward the characteristics combination network model of ResSE_CNN1D, and solve the problem of fitting. ResSE_CNN1D is a one-dimensional time sequence of convolution neural network model, which improves model performance introduced the residual connectionsand SE modules to enhance model performance and performance, while preventing the fitting. The model retains the advantages of the cnn1d model and combines modern CNN improvement technology for classification tasks. ResSE_CNN1D is a one-dimensional time sequence of convolution neural network model, which improves model performance and preventing the over-fitting by introducing the residual connections and SE modules. The model retains the

advantages of the cnn1d model and combines modern CNN improvement technology, which applied to classification tasks.

The innovation of this study is at the following points. Firstly, a lightweight speech emotion recognition model can be proposed to achieve efficient learning of decimal sets and rapid deployment in resource-constrained environments. Secondly, the characteristics of the combination network is used to optimize the classification process of emotional characteristics, and improve the ability of the model to identify complex emotions. Thirdly, the effective regularization and model selection method are proposed for the problem of fitting.

The first part of this article describes the introduction, and the second part shows the combination of the model and analyzes the structure of the model. The third part shows the conclusion of the experiment with the casia data set. Finally, the fourth part is the summary of the full paper.

2. The composition of the model

2.1. Method overview

The structure of this model is shown in figure 1:

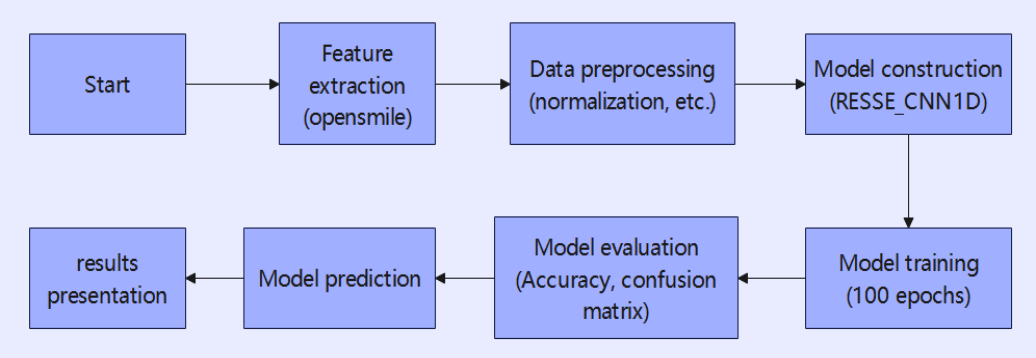

Figure 1. the overall model structure diagram of the experiment

First, the characteristics of the audio species are extracted by the opensmile method[4]. Then the audio is preprocessed, and the audio of the same emotion is grouped together and renamed. The whole data is then trained through the training script, where you can choose to use different models for training, including CNN, RNN, ResSE_CNN1D and other models. After the model is trained, the predictive script can be applied to predict the categories of new audio samples.

2.2. Model structure of ResSE_CNN1D

The ResSE_CNN1D model was utilized in this work. In order to successfully eliminate the issue of gradient disappearance, as well as to add important features and suppress unnecessary features to boost the accuracy of the model, residual and SE modules are added to the deep convolution-based model. The model's general structure is depicted in Figure 2.

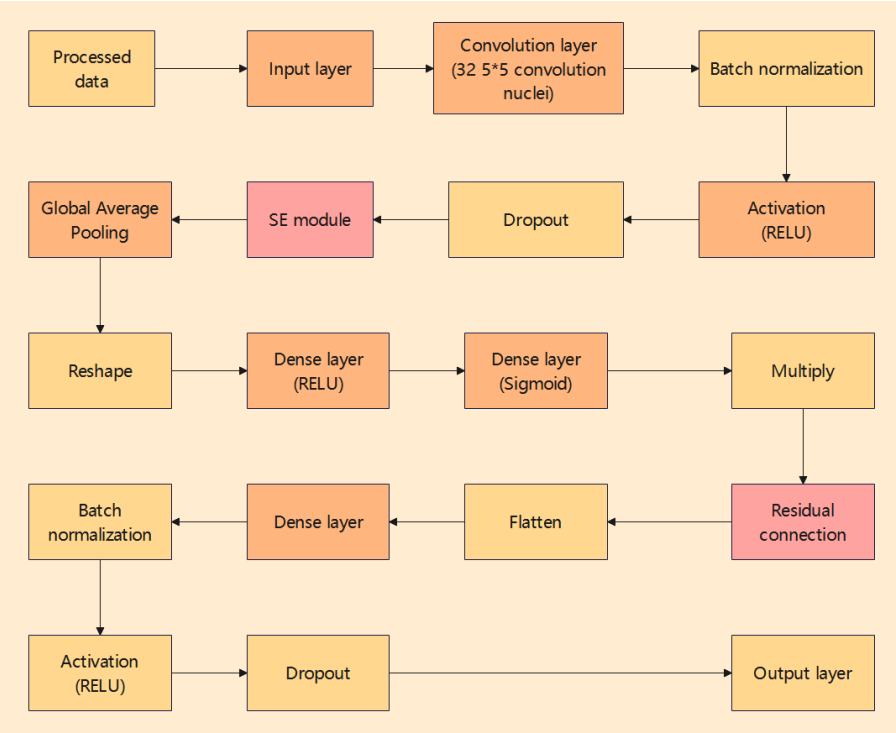

Figure 2. Model structure diagram of ResSE_CNN1D

The ResSE_CNN1D model is improved and extended on the basis of the one-dimensional convolutional neural network model (CNN1D) , aiming to enhance the processing ability of the eigenvalue and eigenvector extracted from opensmile. The following is function of each part

• Input Layer

The input layer is where the model begins. It is responsible for receiving the feature data of the audio processed by opensmile. The direct input of these eigenvalues and eigenvectors also ensures that the model can directly use these data which contain complete information.

• Convolutional Layer

The input layer is followed by a convolutional layer. In order to extract the spatial properties of the input data, each of the 32 5*5 convolution kernels in this layer glides over the input data. The design of multiple small convolution kernels speeds up the overall process and improves the generalization ability of the model. Then immediately following the convolutional layer is the batch normalization layer. By normalizing the variance and mean of each batch of data[5], the batch normalization layer improves the stability of the model and enhances the generalization ability of the model. The next layer is RELU activation layer that imports nonlinear factors, which enhances the nonlinear ability of the model[6], thereby enhancing the expression ability of the model and preventing the occurrence of the gradient disappearance problem. Finally, in order to prevent the occurrence of overfitting problem and lessen the model's reliance on the training set of data, a Dropout layer is introduced, which randomly removes some neurons[7], so as to increase the generalization ability of the model and prevent the occurrence of overfitting problem.

• Squeeze-and-Excitation Module

The SE module, which adds the channel attention mechanism to enhance the weight of significant features and decrease the weight of unnecessary features, is added after the convolutional layer[8]. This module can make the whole model focus more on the processing of the useful parts of the audio, enhance the representation ability of the model, and reduce the amount of calculation.

• Pooling Layer

The features retrieved from the convolutional layer can be further processed by the pooling layer. The global average pooling strategy is employed by the ResSE_CNN1D model in this study to average each feature map into a particular numerical value. In addition to eliminating some noise and repetition, the procedure can combine the audio's emotive elements. This method can achieve the lightweight nature of the model and simplify its structure and amount of parameters

• Fully Connected Layers

After the reshape operation, the output of the pooling layer enters into the fully connected layer. Since every neuron in the fully connected layer is coupled to every other neuron in the previous layer, local features are effectively combined into global information, providing the groundwork for sentiment categorization that comes later.

The RELU function, which maintains the positive eigenvalues and sets the negative eigenvalues to zero, is the activation function of the first completely linked layer. This can enhance the nonlinear ability of the model and help the model converge quickly during the training process. The sigmoid function, which maps the input values between 0 and 1, is the activation function of the second completely connected layer. This helps with the accuracy calculation process in the future.

• Resnet Connection

In order to build a deeper network structure and not cause the problem of gradient disappearance and gradient explosion, this model adds a Resnet connection. In this connection mode, the input layer data skips some intermediate layers and is added to the data processed by these layers, and finally becomes the input of the next layer [8]. By doing this, the vanishing gradient issue is successfully avoided, enabling the model to operate at higher levels.

• Output Layer

Finally, the output data from the residual connection is passed through a fully connected layer by flattening, then through batch normalization, activation layer, Dropout layer, and then into the final output layer. The final classification result is obtained by outputting the predicted probability of each category in the output layer, where the Softmax activation function is employed for classification[6].

3. Experiment

3.1. Referencing datasets

In this experiment, opensmile is used to extract features of audio, and ResSE_CNN1D is used to classify the extracted features. Finally experiments are carried out on the dataset of CASIA[9].

The Chinese Academy of Sciences' Institute of Automation produced the Chinese voice dataset known as CASIA voice Emotion Recognition in order to do research on emotion recognition. It includes 1200 audio recordings of four individuals (two men and two women) expressing the following six emotions: fear, anger, happiness, sadness, neutrality, and surprise.

3.2. Evaluation index

In this experiment, the weighted average precision was used as the test index[10]. The formula for the weighted average precision is as follows.

$$W_A = \frac{\text{Number of correctly identified test samples}}{\text{Total number of test samples}}. \quad (1)$$

The accuracy of the original CNN1D model trained on the dataset of CASIA is about 0.85. After adjusting the initial value and the number of iterations, the highest accuracy of CNN1D was determined to be 0.871.

However, after adding the residual connection and SE module, the accuracy of the whole model can reach 0.900, an improvement of 2.9%.

3.3. Setting hyperparameters

This experiment is implemented by using the TensorFlow-gpu framework, the number of training epochs is 100, each convolution layer has 32 convolution kernels and each convolution kernel size is 5*5. This is because multiple convolution kernels can extract more features and reduce the amount of computation. The batch size is 32, which takes into account the accuracy while ensuring the convergence speed to prevent falling into a local optimal solution. A learning rate of 0.001 is used for this experiment, that is, the step size of 0.001 is used to update the parameters during training. This can lead to a relatively accurate classification at the fastest speed without getting stuck in a local optimal solution. The Dropout rate is set to 0.5 to make the model more robust and thus prevent overfitting. The Hidden_size value is 32, which ensures that there are neither overfitting nor underfitting.

3.4. Analysis of the Experiment

The experiment uses the opensmile toolkit in the process of audio processing. Table 1 is just the feature matrix obtained by opensmile for part of the audio in CASIA.

Table 1. The feature matrix extracted by opensmile

	1	2	3	4	5
1	7.200000e+01	0.000000e+00	5.631022e-01	2.144651e-03	3.507818e-01
2	2.400000e+01	3.000000e+00	7.037972e-01	-2.614775e-04	7.198781e-01
3	2.700000e+01	8.000000e+00	5.141575e-01	-1.616943e-03	6.216841e-01
4	1.360000e+02	2.060000e+00	4.867494e-01	1.432655e-03	3.176961e-01

While obtaining the final results, this experiment added a confusion matrix to visually evaluate the performance of the model, as shown in Figure 3 below. It can be seen from this confusion matrix[11] that the accuracy of the model is high, that is, the classification is more accurate. However, there will be more audio confusion in the recognition of sad and fearful. This is due to the fact that the two emotions are close to each other, which is very difficult to distinguish for machine learning. However, there won't be many mistakes made while identifying a number of other emotions, particularly the group of angry emotions.

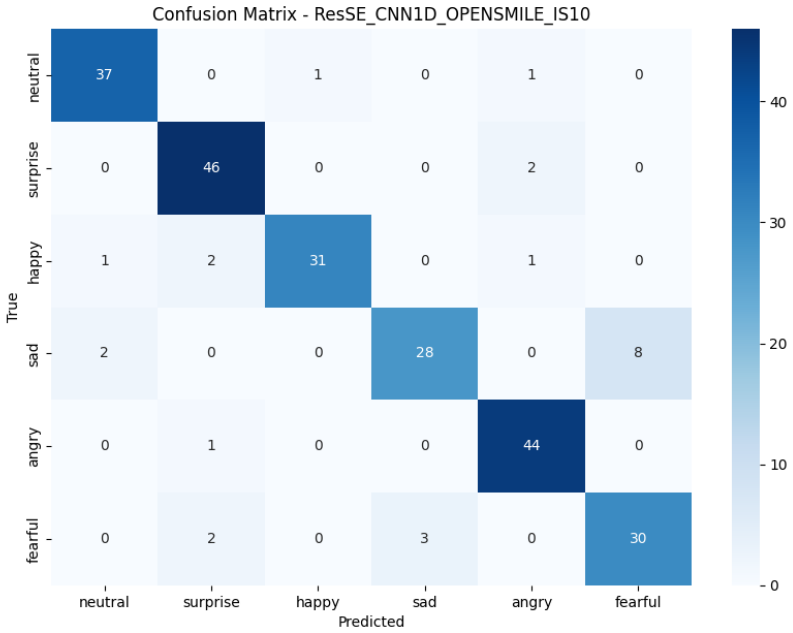

Figure 3. Confusion matrix of ResSE_CNN1D

In order to visualize the results of accuracy during training, this experiment also adds a graph of model accuracy and number of epochs, as shown in Figure 4.

Among them, it can be seen that the test accuracy curve gradually approaches 0.9 after 20 training times, which indicates that the model can obtain a more accurate result with a very small amount of calculation, and realize the goal of lightweight.

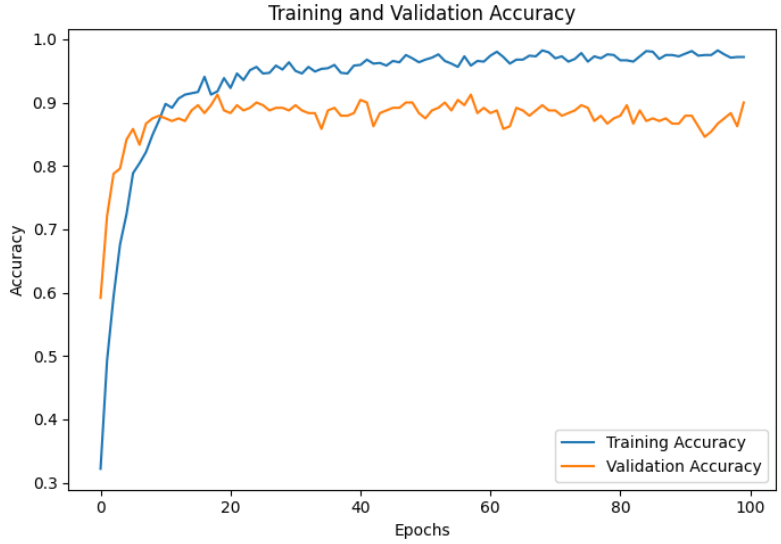

Figure 4. Graph of accuracy versus number of iterations for ResSE_CNN1D model

4. Conclusion

We present a ResSE_CNN1D model in this paper. We introduce Residual Connections and SE module to enhance model performance and representation while preventing overfitting. And the algorithm's efficacy is confirmed using the publicly available CASIA dataset. In the future research, the real-time performance of speech emotion recognition should be considered first, and the real-time

status of speech should be analyzed. Additionally, to be able to identify emotions more rapidly and precisely, the model's generalization rate and accuracy must be further improved.

References

[1] Huang W, Wu Q, Dey N, et al. Adjectives grouping in a dimensionality affective clustering model for fuzzy perceptual evaluation[J]. 2020.

[2] Moschitti A, Pang B, Daelemans W. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)[C] In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[3] Zhao X. XU X. Speech Emotion Recognition Based on Shallow Learning and Deep Learning Models [J]. Computer Applications and Software, 2020,37(12): 108-112,176.

[4] Eyben, F., Wöllmer, M., & Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, pp. 1459-1462. (2010, October).

[5] Ma, Y. and D. Klabjan. "Diminishing Batch Normalization." IEEE Trans Neural Netw Learn Syst 35(5): 6544-6557. (2024)

[6] Sharma, S., Sharma, S. and Athaiya, A. Activation functions in neural networks. Towards Data Sci, 6(12), pp.310-316.(2017)

[7] Park, S. and Kwak, N. Analysis on the dropout effect in convolutional neural networks. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 189-204). Springer International Publishing.(2017)

[8] Hu, J., Shen, L. and Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).(2018)

[9] J. H. Tao, F. Z. Liu, M. Zhang and H. B. Jia, "Design of speech corpus for mandarin text to speech", Proc. Blizzard Challenge Workshop, pp. 1, 2008.

[10] Zhang J. Zhang S. Yan Q. et al. Emotion Recognition method based on speech rhythm difference. Computer Science,2024,51(4):262-269. DOI:10.11896/jsjkx.230200063.

[11] Deng, X., Liu, Q., Deng, Y. and Mahadevan, S. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences, 340, pp.250-261.(2016)