2021 International Conference on Communications, Information System and Computer Engineering (CISCE) | 978-1-6654-0352-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CISCE52179.2021.9445906

2021 IEEE 3rd International Conference on Communications, Information System and Computer Engineering (CISCE 2021)

Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network

Siqi Han
Beijing University of Posts and Telecommunications
Beijing, China
Email: duoerxs@163.com

Feng Leng
Harbin Medical University
Daqing City, Heilongjiang Province, China

Zitong Jin
North China Electric Power University
Baoding Hebei Province, China

***Abstract*—As a challenging pattern recognition task, speech emotion recognition has attracted more and more attention in recent years and is widely used in medical, Affective Computing, and other fields. In this paper, we proposed a parallel network of ResNet-CNN-Transformer Encoder. The Res-Net is used to alleviate the problems caused by the deepening of the network. The CNN calculates the fewer parameters to increase the fitting expression ability of the network. Due to the traditional recurrent neural network, with a long-term dependence on the feature extraction of speech and text sequences and sequence attributes not capturing long-distance features, the multi attention mechanism of the transformer coding layer is used to parallelize the sequence, improve the processing speed and extract the emotional semantic information in the sequence. Experiments are carried out on the RAVDESS dataset. Our results demonstrate the effectiveness of the proposed method and make a significant improvement compared with the previous results.**

Kewords-component：Residual neural network 、CNN、Emotion recognition、 Transformer-Encoder

I. INTRODUCTION

Human emotions are important in communicating and making decisions with others. Identifying emotions is significant in intelligent human-computer interaction (HCI) applications, such as virtual reality, video games, and education systems. In the medical field, detected patient mood can be used to indicate certain functional disorders, such as major depression. As the most direct means of communication, speech contains a wealth of emotional information, and the changes in people's emotions can be reflected through the characteristics of speech. Speech emotion recognition converts the input speech signals containing emotional information into readable physical features. It extracts the speech features related to emotional expression, which constructs the emotion recognition classifier for testing and training. Finally, it outputs the classification results of emotion recognition. Han et al. [1] used Deep Neural Networks (DNN) to extract deeper features from original data and verified the effectiveness of Deep Neural Networks for speech emotion recognition. However, this structure has some limitations for long-distance feature extraction. Mustaqeem et al. [2] proposed a new SER framework based on a radial basis function network, which improves the accuracy of speech emotion recognition for clustering similarity measurement of key sequence selection. Liu et al. [3] proposed a method of speech feature extraction based on formant feature extraction and phoneme type, which significantly reduced the time required for speech recognition.

Deep Neural Network is a powerful tool for speech recognition. Its advantage lies in the ability to extract the main features of speech materials independently through models. In the process of speech recognition [4], the feature selection technology based on correlation is applied to the extracted features to conduct emotion classification through the most discriminative and appropriate emotion features.

This work is based on the open dataset RAVDESS [5]. Through the application of the Res-Net, the data output of the first several layers is directly introduced into the input part of the later data layer by skipping the multiple layers to alleviate the problems about the gradient disappearance caused by the deepening of the network, such as the decrease in the accuracy of network optimization and the decrease in learning efficiency. Using multiple convolutional layers with smaller convolutional cores, it is used to replace one convolutional layer with larger convolutional cores. On the one hand, it can reduce parameters. On the other hand, it is equivalent to carrying out more nonlinear mappings and increasing the network's fitting expression ability. Because the traditional Recurrent Neural Networks cannot capture characteristics for long distances, the Transformer using the self-attention mechanism to realize fast parallel, which directly accesses global information to solve the sequence distance limit and makes the best of the emotional semantic information extraction in the sequence.

Meanwhile, it can also improve the processing speed[6]. In this paper, the network depth and generalization ability are increased by constructing the parallel network of Res Net-CNN-Transformer Encoder. At the same time, the GELU activation function is used to avoid the problem of gradient disappearance.

The rest of this paper is organized as follows: The dataset and feature extraction are presented in Section II. Section III explains the proposed model in detail. The experiments and results are shown in Section IV. The proposed work is concluded in Section V.

II. DATASET AND FEATURE EXTRACTION

A. Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is selected as the dataset for our model

978-0-7381-1215-2/21/$31.00 ©2021 IEEE

due to its high quality. The dataset contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgusted expressions. The total number of samples is 1440.

Since the RAVDESS is small and clean, it can easily lead to overfitting. In addition, due to the use of ResNet, our model parameters have increased a lot compared with the general CNN, which makes the model more prone to overfitting. In order to overcome these problems, data enhancement by adding white noise is performed. White noise is Gaussian because the noise vector will be sampled from a normal distribution and have zero (zero-mean) time. A whitening transformation processes it. After that, the noise will add power to the audio signal uniformly across the frequency distribution.

Two additional samples containing white noise are generated for each sample, and used these three samples together as the dataset. Now the overall number of samples is 4320.

B. *Feature Extraction*

Feature extraction plays a vital role in achieving good training results. Mel spectrograms and MFCCs are widely utilized to extract features in speech emotion recognition and classification tasks. MFCC is a mathematical method that transforms the power spectrum of an audio signal to a small number of coefficients representing the power of the audio signal in a frequency region (a region of pitch) taken w.r.t. time. MFCCs are commonly derived as five steps: First, take the Fourier transform (a windowed excerpt of) a signal. Then map the powers of the spectrum obtained above onto the mel scale, using overlapping triangular windows. After that, take the logs of the powers at each of the mel frequencies. Next, take the discrete cosine transform of the list of mel log powers as if it were a signal. Finally, the MFCCs are the amplitudes of the resulting spectrum. We only utilize mel spectrograms and Mel-frequency cepstral coefficients (MFCCs) to extract features in the datasets. The reason is that after experiments, we find mel spectrograms + MFCCs alone provide the best accuracy in this model with training considerations in mind. So we do not want extra complexity in a highly parameterized deep neural network especially using ResNet.

C. *Model*

In the proposed method, we build a parallel model that includes three parts: the Residual Neural Network (ResNet), the Convolutional Neural Network, and the Transformer part.

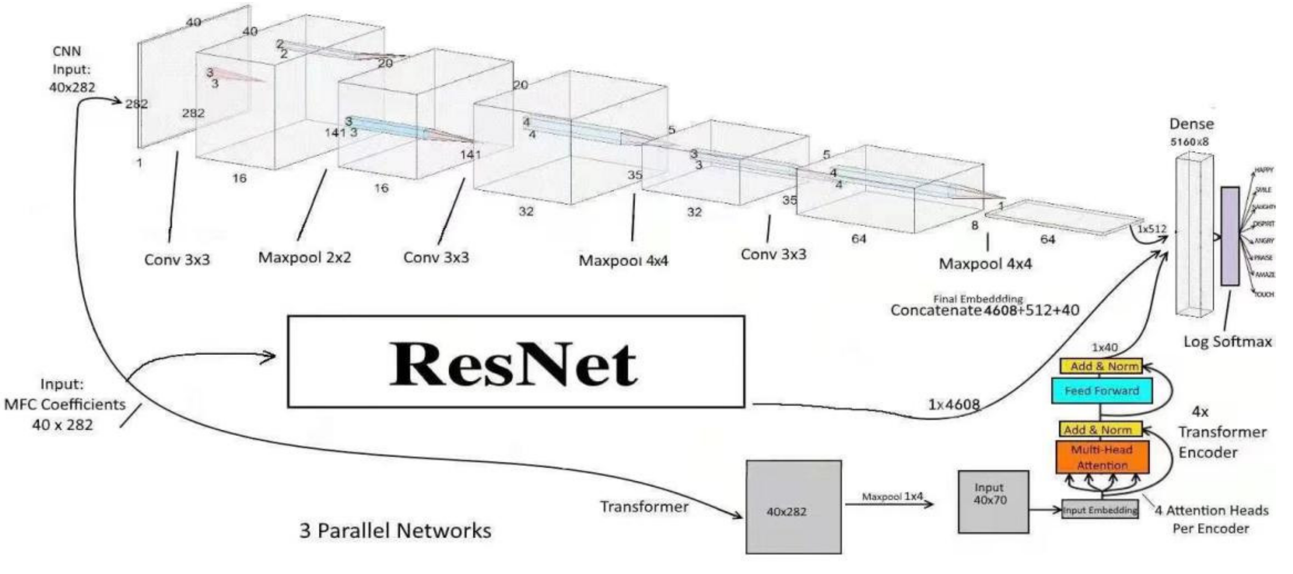


Figure 1. Architecture of our model

III. Residual Neural Network (ResNet)

A. *Residual Neural Network (ResNet)*

Residual Neural Network (ResNet) is an artificial neural network (ANN) built on the known structure of the pyramidal cells of the cerebral cortex. Residual neural networks do this by using shortcuts to skip connections or skip specific layers. In previous research, typical ResNet models are implemented with multi-layers (mostly double- or triple-layer) skips that include nonlinearities (ReLU) and batch normalization.

One motivation for skipping over layers is to avoid the problem of vanishing gradients by reusing activations from a previous layer until the adjacent layer learns its weights. Skipping simplifies the network, using fewer layers in the initial training phase. Since the number of propagation layers is reduced, the learning speed can be accelerated by reducing the vanishing gradient. Then, the network will gradually restore the skipped layers as it learns the feature space. When all levels are expanded at the end of the training, it will be closer to the manifold, so the learning speed is faster.

In our model, we establish a structure similar to resnet18, using four layers of ResNet Blocks. Meanwhile, the kernel size, stride, and padding in each layer are the same as resnet18. To prevent overfitting, we added the dropout blocks between each layer. In addition, through experiments, we found that using GELU as the activation function in this task can achieve higher accuracy. The Gaussian Error Linear Units(GELU) function is as follows:

$$\mathrm{GELU}(x) = x * \Phi(x) \quad (1)$$

Where $\Phi(x)$ is the Cumulative Distribution Function for Gaussian Distribution, the most notable feature of the GELU activation function is nonlinearity and random regularizers that depend on input data distribution in an activation function expression. Unlike the previous way that dropout specifies random probability values or ReLU masks based on the positive and negative input values, GELU performs random regularization based on the probability that the current input is greater than the rest of the inputs. Although GELU is more used in NLP tasks, we still use GELU in ResNet to improve the classification effect of the model.

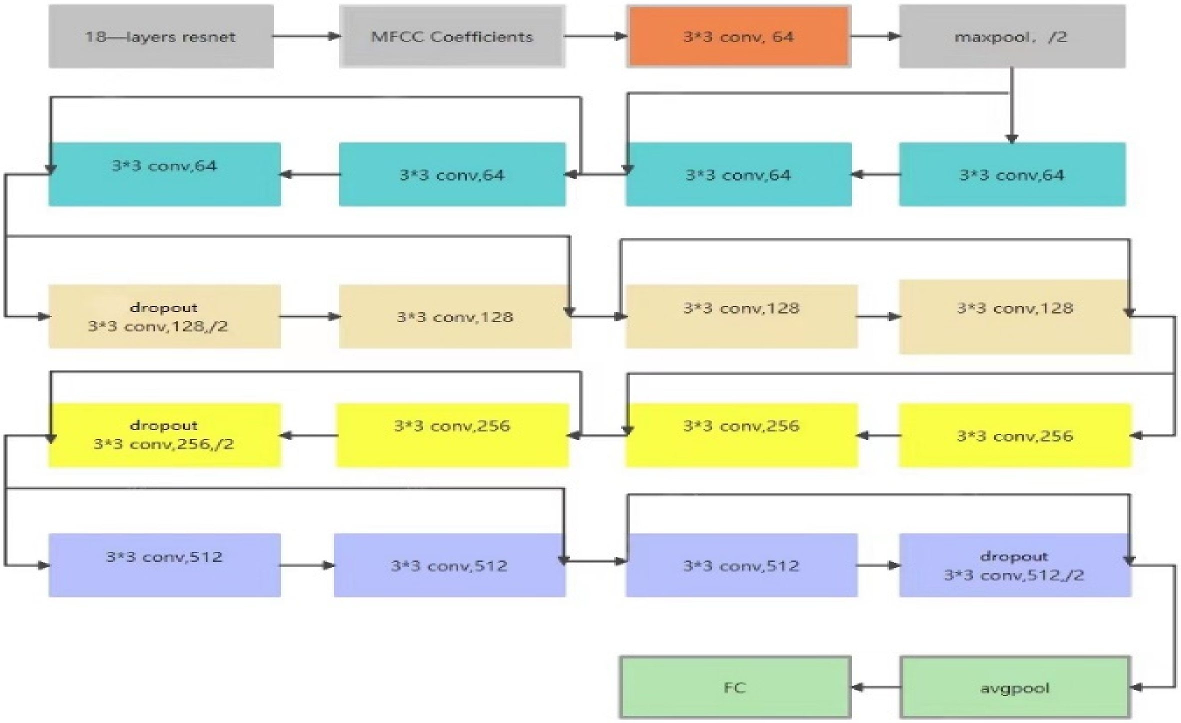


Figure 2. Architecture of ResNet

B. Convolutional Neural Network (CNN)

CNN's have image and video recognition applications, brain-computer interfaces, and speech emotion recognition and achieve impressive results. Our model utilizes 3 convolutional layers, 3 max pool layers as a feature extractor to extract features in MFCCs. The first layer of our CNN model receives 40 × 282 arrays as input with a 3 × 3 kernel size. Then batch normalization and activation function GELU are applied. After that, the max pool layer is utilized to speed up calculation and prevent overfitting. The kernel size and stride of the max pool layer are both set to 2.

At last, we use dropout with the rate of 0.5. We repeated this structure three times in our CNN model. The only difference is that the kernel size and stride in the max pool layers of the last two times are both set to 4.

C. Transformer Encoder

Considering the continuity of speech in time, we use the transformer's multi-head self-attention layers to establish our third model. The purpose of designing the Transformer is that the network will learn to predict the frequency distribution of different emotions according to the overall structure of the MFCC of each emotion.

The first layer is a max pool layer which receives 40 × 282 arrays as input. After that 4 multi-head self-attention layers are applied with activation function GELU and dropout (with the rate of 0.4), the shape of arrays becomes 40 × 70. Finally, we calculate the mean value on the second dimension to make it a one-dimensional vector (40 × 1).

The reason for not using RNN or LSTM to learn the order of the spectrogram of each emotion is that they only learn to predict the frequency change based on adjacent time steps. In contrast, the transformer's multi-headed self-attention layer allows the network to look at multiple previous steps and time steps when predicting the next step. It makes sense because emotions exist throughout the sequence, not just over a time step.

D. Concatenate Three Parts

To integrate the three models, we flatten the results of ResNet and CNN into a one-dimensional vector, respectively, and concatenate them with the results of Transformer Encoder. Then a fully-connection neural network and softmax are applied sequentially to calculate the probabilities of 8 emotion status.

IV. Experiment and Results

A. Experiment

We choose CrossEntropyLoss as loss function and stochastic gradient descent (SGD) as the optimizer method in our experiment. We divide the dataset into the training set, validation set, and test set according to the ratio of 80:10:10.

B. Results

We ran the model five times on the dataset and computed the average classification accuracy on test sets. The accuracy of emotion classification is recorded as 80.89% after 500 epochs.

The current human accuracy rate for this dataset has been reported as 67% [9], which indicates that the classification of emotions for this dataset is not a simple and straightforward task, even for human beings. Figure 3 illustrates the comparison of our model with the previous works on speech emotion recognition using the RAVDESS dataset. According to the results, the proposed framework considerably outperforms the models of Shegokar and Sircar [7], Zeng et al. [8], and human accuracy [9].

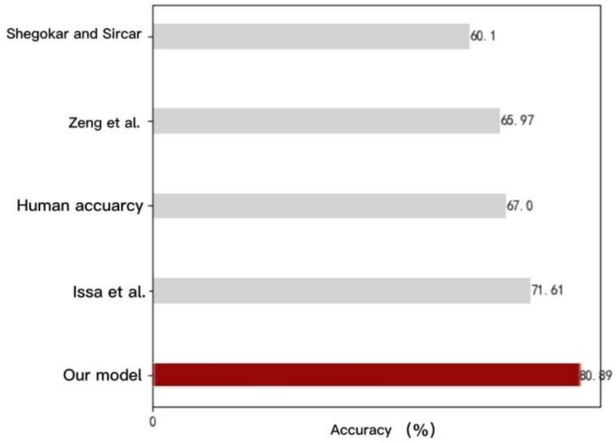
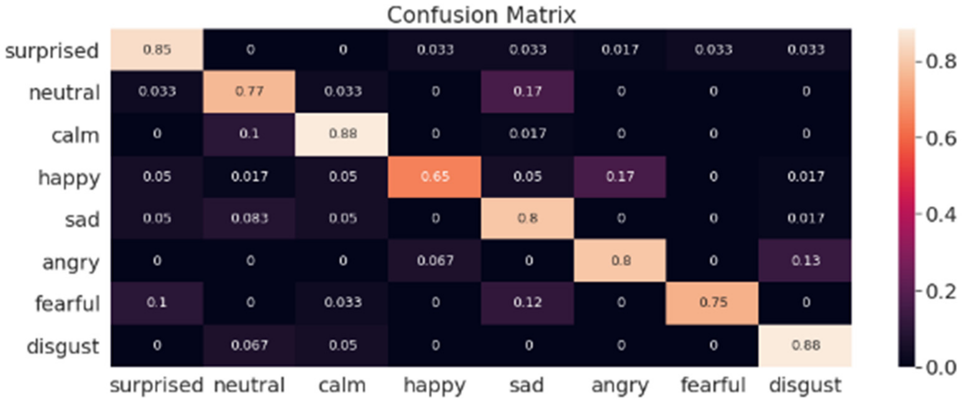


Figure 3. Comparison with the previous works

We calculated the confusion matrix, which is as follows:



	surprised	neutral	calm	happy	sad	angry	fearful	disgust
surprised	0.85	0	0	0.033	0.033	0.017	0.033	0.033
neutral	0.033	0.77	0.033	0	0.17	0	0	0
calm	0	0.1	0.88	0	0.017	0	0	0
happy	0.05	0.017	0.05	0.65	0.05	0.17	0	0.017
sad	0.05	0.083	0.05	0	0.8	0	0	0.017
angry	0	0	0	0.067	0	0.8	0	0.13
fearful	0.1	0	0.033	0	0.12	0	0.75	0
disgust	0	0.067	0.05	0	0	0	0	0.88

Figure 4. Confusion matrix

It can be seen from the figure that the best classification effect is on “disgust” and “calm.” However, the classification accuracy of “happy” is low.

V. Discussion and Conclusion

Unlike image classification and general language emotion classification tasks, speech emotion classification is a complex task in which feature extraction and the classification model are the two most important parts. This paper concatenates ResNet, CNN, and Transformer Encoder to propose a new framework (parallel model) for speech emotion recognition on the RAVDESS dataset. We use GELU as an activation function and achieve 80.89% accuracy, which is higher than current results. In the subsequent research, we believe that the following things can further improve classification accuracy. Trying other feature extraction methods, such as calculating Power Spectral Density (PSD) or using Singular Value Decomposition (SVD), may improve the effect of feature extraction. Besides, improving the neural network structure, such as increasing the number of layers of ResNet, may also improve the classification effect. Finally, Increasing the number of ensemble models may also help the entire model to complete the classification.

Reference

[1] Han K, Y u D, and TashevI. Speech emotion recognition using deep neural network and extreme learning machine[C]//in Proc. 15th Annu. Conf. Int. Speech Commun. Assoc, 2014: 223–227.

[2] MustaqeemY, Sajjad M and Kwon S. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep Bi-LSTM [J]. IEEE Access, 2020, 8:79861-79875.

[3] Liu Zhen-Tao et al. Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence[J]. Information Sciences, 2021, 563 : 309-325.

[4] Farooq Misbah et al. Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network.[J]. Sensors (Basel, Switzerland), 2020, 20(21)

[5] SR. Livingstone, F.A. Russo, 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 (2018) e0196391.

[6] Lee, S.S. Narayanan, Iemocap: interactive emotional dyadic motion capture database,Lang. Resour. Eval. 42 (2008) 335.

[7] P. Shegokar, P. Sircar, Continuous wavelet transform based speech emotion recognition, in: 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), IEEE, 2016, pp. 1–8.

[8] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimed. Tools Appl. (2017) 1–18.

[9] S.R. Livingstone, F.A. Russo, 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 (2018) e0196391.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.

[11] 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.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[13] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[14] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929-1958.

[15] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International conference on machine learning. PMLR, 2015: 448-456.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[17] Issa D, Demirci M F, Yazici A. Speech emotion recognition with deep convolutional neural networks[J]. Biomedical Signal Processing and Control, 2020, 59: 101894.