Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

End to End Speech Recognition Based on ResNet-BLSTM

Cunzhu Zou, Jiping Luo, Chuangjie Huang

Information Science and Technology College, Dalian Maritime University, Dalian, China

1462920155@qq.com

Abstract: In the end-to-end speech recognition model based on deep learning, the input of the model uses fixed-length speech frame, which results in the loss of time-domain information and part of high-frequency information, resulting in low recognition rate and poor robustness. Because of the above problems, this paper proposes a model based on the combination of residual network and bidirectional long and short-term memory networks. The model uses spectrogram as input, designs parallel convolution layer in residual network, extracts features of different scales, and then performs feature fusion. Finally, it implements an end-to-end speech recognition model by connecting time series classification method. The experimental results show that the WRE error rate of the proposed model on AISHELL-1 speech set is 2.52% lower than that of the traditional end-to-end model, and its robustness is better.
Keywords: ResNet; BLSTM; Parallel convolutional layer; Connectionist temporal classification.

---

1. INTRODUCTION

Automatic speech recognition (ASR) is a complicated task. The system is based on a strictly designed processing flow, including input features, acoustic model, language model, and hidden Markov models (HMM) [1]. ASR's long-term goal is to be able to deal with people's speech effectively in different environments, which is quite challenging for the traditional gmm-hmm hybrid model. An excellent acoustic model should have the ability to effectively simulate various acoustic changes in the speech signal, to obtain the robustness under different speech and environmental conditions. In speech recognition model, feature extraction and classifier are usually considered as two independent problems. Firstly, the feature extraction method is designed, and then the classifier performance is optimized based on the extracted features. There are two disadvantages of this method [2-3]: one is that the feature extraction method of artificial design needs careful design and much experimental verification; the other is that the feature of artificial design is not guaranteed to be optimal for the current classification task. DNN and its variants can simultaneously perform feature extraction and classification tasks. As shown in literature [4], the lower level of DNN can extract features adaptive to speakers, and the higher level of DNN can improve the differentiation of different categories, to improve the final classification effect. One of the essential reasons why DNN is superior to Gaussian mixture model (GMM) is the joint optimization of feature normalization and classification tasks.

In recent years, with the development of deep learning technology, revolutionary neural (CNN) and long short term memory (LSTM) have achieved a better recognition rate than the traditional speech recognition technology in speech recognition technology. However, the input characteristic of its network structure is usually Mel frequency cepstral coefficients (MFCC). This characteristic is proposed based on the characteristics of human hearing. It has a non-linear relationship with the frequency, but it is easy to cause the loss of information in the high-frequency region. Moreover, the traditional speech characteristics must adopt a very large frameshift in order to consider the calculation amount, which undoubtedly causes the signal in the time domain. The loss of breath is more prominent when the speaker speaks faster. Therefore, for speech recognition task, it is not guaranteed to be the optimal feature.

Furthermore, in order to solve these problems, literature [6] studies how to extract better features than MFCC from the original speech waveform as the input of the neural network by using neural network training methods. This method does not need any manual intervention, and the parameters of the network are determined by training data and objective function. In reference [7], a layer of time-domain convolution is used as the feature extractor, and the hybrid neural network of revolution LSTM deep neural network (CLDNN) is used. For the first time, the model recognition performance with the original speech waveform as the input is achieved with the MFCC feature as the input. However, because they use only a few convolution layers, generally only 1-2 layers, and only use CNN as a feature extractor, such convolution network structure expression ability is very limited. In order to solve this problem, in the reference [8], a model combining multi time-frequency resolution convolution network and feed-forward sequential memory network (FSMN) with memory module is adopted, which improves speech recognition performance and training speed. However, using more FSMN layers will make the gradient disappear and lead to the instability of training.

To solve the above problems, this paper proposes a model structure based on the combination of RESNET [9] and BLSTM [10] to improve the recognition rate and system robustness of the end-to-end model. By using the residual network, the speech spectrum of the whole speech is directly input, which is faster than other speech recognition models with traditional speech features as input. Secondly, from the perspective of model structure, RESNET in this paper is different from CNN in the traditional end-to-end system. It draws lessons from the approach in image recognition, by transforming voice into an image as input, then taking time and frequency as two dimensions of the image, and then through some combination of convolution layer and pooling layer, CNN's expression ability is greatly enhanced. Secondly, after RESNET is connected to BLSTM, the context information of speech signals can be learned through the network to improve the recognition rate of end-to-end model.

The contribution of this paper is three-fold:

We propose CNN+BLSTM+CTC, an end-to-end ASR model using both CNN and BLSTM. It combines CNN layer’s ability of learning local features and BLSTM layer’s ability of learning history and future contextual features, enabling the CNN+BLSTM+CTC to model audio signals effectively and make precise recognition.

We use neither in-house training data no external language model in this paper. All the training, development, testing data we used come from dataset AISHELL-1, which can be freely acquired. This makes our results comparable for other researchers.

We carry out comprehensive experiments to verify our design ideas. Experiments results show that our CNN+BLSTM+CTC makes effective speech recognition.

2. RELATED WORKS

Commonly, for an input utterance, conventional state-of-art ASR systems use HMM-based acoustic model to extract acoustic features, use GMM-based pronunciation model to map acoustic features to sub-phonetic states and use pronunciation lexicon to map sub-phonetic states to a sequence of words. Finally, the word sequence is rescored by external language model to generate a reasonable sentence. Models working in such way have many disadvantages.

Building such an ASR system is a very tough work. Firstly, there are many modules in such a system such as acoustic model, language model, to name but a few. Secondly, different domain knowledge and expert engineering work are needed to design these different modules. For example, a linguistics expert may be needed to design the language model.

Training a good-performing model is very hard. Since different modules are designed based on different hypotheses, they need different expertise for training. What makes things worse, each of them has its own optimizing objectives, which may be different from each other and even different from the global optimizing objective. All these together make it difficult to train a good-performing model.

These models are awkward to fine tune. As they contain many modules, when we want to adapt them to recognize speeches in a new scenario, most of these modules must be retrained from scratch, which will cost a lot of time and effort.

Structure of such models is inflexible. Modules contained in a conventional model and the structure between these modules are almost fixed. It is hard to add/delete/change a module or reorganize their structure. Thus, it is difficult to introduce new developed technologies such as deep learning into these models.

These models need high-quality dataset for training. The training data must be aligned, which means that every input frame must have a corresponding label. Building such a dataset takes masses of time, effort and domain knowledge, and must be very careful. As a result, it is almost impossible to build a large-scale dataset.

Recently, researchers have been working on end-to-end ASR methods to overcome these disadvantages of conventional ASR.

End-to-end ASR is a kind of sequence-to-sequence model. In contrast to conventional ASR that contains many modules and derives the final result from several intermediate states, end-to-end ASR directly maps input acoustic signals to graphemes such as characters or words. It subsumes most modules into a DNN and use an overall training objective function to optimizes the criteria that related to the final evaluation criterion we really concern about (in most cases, it is the Word Error Rate, WER). However, in conventional ASR, every module has its own objective function, which is indirectly related to the final evaluation criterion.

3. DCNN MODEL

As a feature extractor of speech spectrum, CNN includes the convolution layer and pooling layer. In the last layer of convolution network, Max pooling is used to sample the output features in a fixed size so that most of the feature information is retained while the parameters are reduced and the performance is improved. Because CNN is good at modeling the local structure of the input, and convolution operation of convolution layer on the spectrum, its convolution weights are all shared, compared with DNN network, it can greatly improve the training speed. DCNN model is shown in Figure 1:

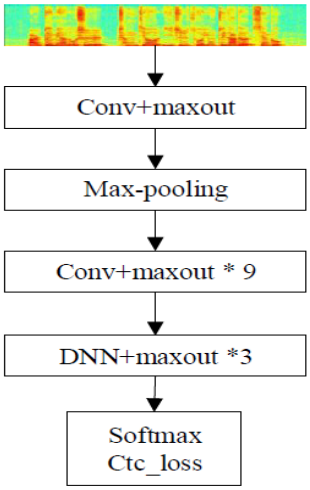


Figure 1. DCNN model.

The major building block of Deep Speech is a recurrent neural network that has been trained to ingest speech spectrograms and generate English text transcriptions. The purpose of the RNN is to convert an input sequence into a sequence of character probabilities for the transcription.

The RNN has five layers of hidden units, with the first three layers not being recurrent. At each time step, the non-recurrent layers work on independent data. The fourth layer is a bi-directional recurrent layer with two sets of hidden units. One set has forward recurrence while the other has backward recurrence. After prediction, Connectionist Temporal Classification (CTC) loss is computed to measure the prediction error. Training is done using Nesterov's Accelerated gradient method.

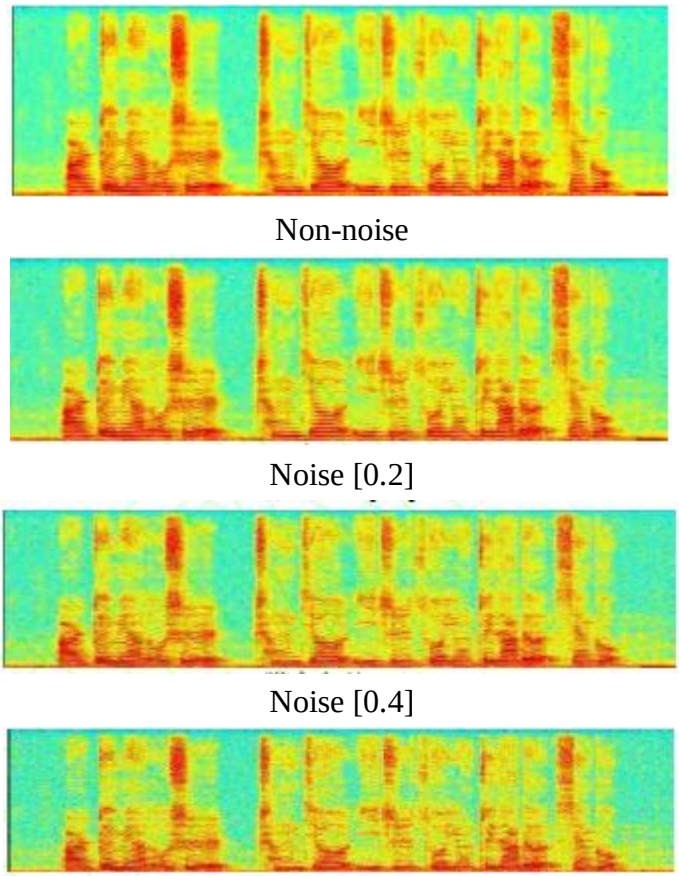
4. RESNET-BLSTM MODEL

With the increase of the number of layers in a single stacked CNN network, the gradient disappears, and the network performance deteriorates. Therefore, this paper introduces residual module and parallel convolution layer in the original CNN, and accesses the BLSTM layer, and proposes ResNet-BLSTM model. The model consists of two modules, ResNet module, and BLSTM module. ResNet is able to extract local features from the spectrum, and then BLSTM is used to model the features in context.

4.1 Model input

Because the time domain analysis cannot directly reflect the frequency characteristics of speech signals, and the frequency domain analysis cannot represent the relationship between speech signals and time, so reference [11]. It combines the advantages of the spectrum chart and time-domain waveform chart. It can display the change of speech signal spectrum with time intuitively. It is a dynamic spectrum. It can be seen intuitively from the spectrogram that at any given time, different

frequency components have different colors, so the spectrum values are different. In this paper, the original speech signal is converted into an N *N three-channel spectrogram, and three sets of Gauss white noise points with different variances [0.2, 0.4, 0.6] are added to the original spectrogram to improve the robustness of the model. The original spectrum and noisy spectrum are shown in Figure 2:



Non-noise

Noise [0.2]

Noise [0.4]

Noise [0.6]

Figure 2 Original spectrum and noisy spectrum

4.2 ResNet module

4.2.1 Residual Network

The residual module is introduced into the stacked CNN network. The basic principle of the residual module is shown in Figure 3:

4.2.2ResNet module design

The convolution neural network is mainly composed of the convolution layer and pooling layer. The more convolution layer, the deeper network depth. Although the deep network can extract more abundant features, it will also lead to the complexity of the model and the problem of too much calculation.

ResNet module includes an 8-layer volume layer, a 2-layer pool layer and a set of parallel volume layers. Because this paper considers the following two aspects when designing the size of convolution core: 1. Spectrogram contains the characteristics of speech signals; 2. Excessive convolution core is easy to increase the amount of computation, which is not conducive to increasing the depth of the network. Therefore, the convolution kernel sizes selected in this paper are 3 * 3, 5 * 5 and 1 * 1, respectively.

At the same time, in order to solve the problem of a low recognition rate caused by speech speed and improve the robustness of the model, a parallel convolution group is designed in this paper. The number of convolution kernels of each group in this layer is 128, but the size is not the same. The purpose is to extract features of different scales. Finally, three sets of output and the output of the upper layer are fused so that the model can adapt to the speaker's speed.

Since most of the information of the speech signal is reflected in the energy spectrum, the highest pooling layer is used in the last layer to retain the texture features in the spectrum.

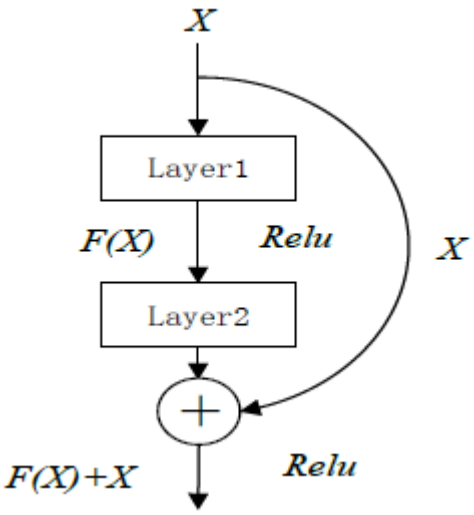


Figure 3 Residual network structure

4.3 BLSTM module

4.3.1 BLSTM network

Due to the long-time characteristic of the speech signal, DNN often uses fixed speech frame as input when processing speech signals, and can't use the information between the front and back speech frames. LSTM [12] can solve these problems well. Figure 4 shows the basic LSTM structure.

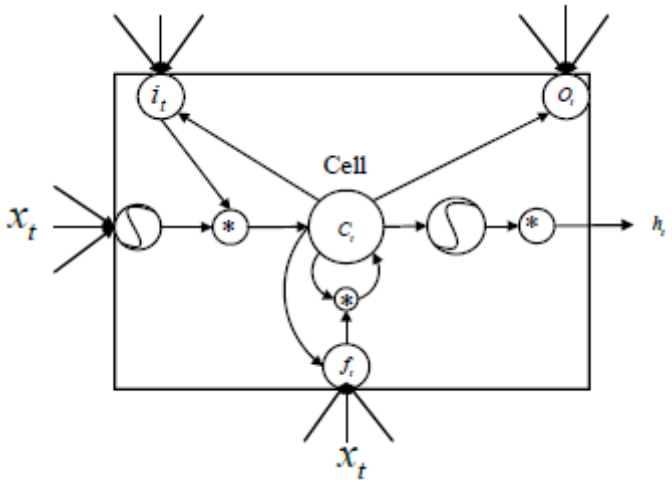


Figure 5 LSTM network structure

BLSTM acts on a forward and backward LSTM network for each training sequence and connects the same output layer so that the network can fully learn the context information of the sequence.

5. EXPERIMENTAL STEPS

5.1 Experimental data

The voice data used in this paper is the open-source voice data set air shell-1 [16], which is recorded in real environment, including multiple participants from different regions of China with different sounds. The environment is in a quiet room, using three different devices for recording: high fidelity microphone (44.1KHz, 16bit); Android system, mobile phone (16KHz, 16bit); IOS system mobile phone (16KHz, 16bit). In this paper, we use 16KHz speech, often 178h. The validation set is 10h, and the test set is 5h. This paper will verify the performance of the model on the verification set and the test set respectively.

5.2 Model training and optimization

In the process of optimizing the model, the initial learning rate used in this paper is 0.01, the optimizer is Adam, and the mechanism of degenerate learning rate is used, that is, every m cycles, the learning rate automatically drops by 0.95. The setting range of the initial weight value is [- 0.1, 0.1], and the initial offset value is 0.1. At the same time, in order to prevent overfitting, dropout method is used in this paper. The initial dropout value of each layer of network is set to 0.95. In the process of calculating the loss value, L2 regularization is used to improve the fitting. The loss function uses CTC - loss, and the specific formula is as follows:

$$L(S) = -\sum_{(X,Y)\in S} \ln p_r\,(Y|X)$$

5.3 Evaluation index parameters

The evaluation index used in this experiment is the word error rate (WER). The purpose is to test the accuracy of recognition results. The specific formula is as follows:

$$WER = 100 \times \frac{S + D + 1}{N} \times \%$$

$$Accuracy = 100 - WER$$

6. ANALYSIS OF EXPERIMENTAL RESULTS

6.1 Experimental comparison of model parameters

To validate this model, the comparison of the three-end recognition model: (1) BLSTM + DNN + CTC model [17]; (2) CNN + LSTM + DNN + CTC [18] model; (3) CNNmaxout + DNNmaxout + CTC [11] model. DETAILED model structure parameters in Table 1:

Table 1 Model Network Parameters 1-3

Model structure	Network Parameters
Model 1	5*DNN+ 3*BLSTM+2*DNN
Model 2	3*CNN+3*LSTM+4*DNN
Model 3	10*CNNmaxout+3*DNNmaxout

MFCC feature 1 wherein the parameters for the 26-dimensional model, the input current plus the front and rear frame 9, a total of 494 parameters dimension. 2 is a model filter characteristic set of features (Filter bank, Frank), the MFCC speech feature is the final step of removing the discrete cosine transform is obtained, as compared with MFCC feature, it retains more of the original speech data. 3 input models and this model are spectrograms.

6.2 Performance model to explore

6.2.1 Model Performance Comparison

In order to study the performance of different models, this paper was first trained on the training set of four models, save four kinds seven times the number of iterations of the model, and then do the validation set verification, to find out the optimal number of iterations, and finally on the test set do performance testing. FIG 7 is a model on the effect of each training set.

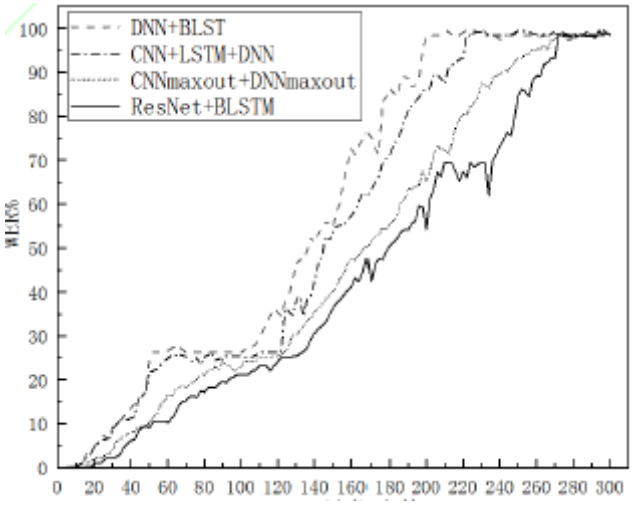


Figure 7 Number of epoch iterations of the training set

It can be derived from Figure. 7, Model 1 and Model 2, respectively, at the iteration to about 200 times, to achieve the best performance of the model.

The use of 3 spectrograms as a model and this model requires the input of iterations to around 260 times to reach optimum performance. This is because this model is the model 3 and convolutional neural network learning feature information through the spectrogram; therefore a large number of iterations are needed. With the increase in the number of iterations, each model tends to fit the training data state.

7. CONCLUSION

This paper presents an end-to-end speech recognition model, which is composed of ResNet and BLSTM and uses a spectrogram as input. Compared with other end-to-end models, the overall performance of this model is better than other models. But the disadvantage of this model is that it is not easy to train continuous long speech.

After that, the next step is to optimize the model, train, and test the noisy speech set.

ACKNOWLEDGMENTS

This work is supported by the innovation and entrepreneurship training program at Dalian Maritime University.

REFERENCES

[1] Petridis S, Li Z, Pantic M. End-to-end visual speech recognition with LSTMS[C]// IEEE International Conference on Acoustics. 2017.

[2]Petridis S, Wang Y, Ma P, et al. End-to-End Visual Speech Recognition for Small-Scale Datasets[J]. 2019.

[3]Wöllmer M, Schuller B, Rigoll G. A novel bottleneck-BLSTM front-end for feature-level context modeling in conversational speech recognition[C]// Automatic Speech Recognition & Understanding. 2011.

[4]Solecasals J. Advances in Nonlinear Speech Processing: International Conference on Nonlinear Speech Processing, NOLISP 2009, Vic, Spain, June 25-27, 2009, Revised Selected Papers[J]. Lecture Notes in Computer Science, 2010, 4885(11):920-921.

[5]Cohen D, Croft W B. End to End Long Short Term Memory Networks for Non-Factoid Question Answering[J]. 2016.
[6]Fung I, Mak B. END-TO-END LOW-RESOURCE LIP-READING WITH MAXOUT CNN AND LSTM[J]. 2018:2511-2515.
[7]Tao Z, Chen M, Jie Y, et al. Attention-Based Natural Language Person Retrieval[C]// Computer Vision & Pattern Recognition Workshops. 2017.
[8]Cai W, Cai D, Huang S, et al. Utterance-level end-to-end language identification using attention-based CNN-BLSTM[J]. 2019.
[9]Xing Y, Liang S, Sui L, et al. DNNVM : End-to-End Compiler Leveraging Heterogeneous Optimizations on FPGA-based CNN Accelerators[J]. 2019.
[10]Chen H, Wang Y, Xu C, et al. DAFL: Data-Free Learning of Student Networks[J]. 2019.
[11]Miao Y, Gowayyed M, Metze F. EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding[J]. 2015.
[12]Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]// IEEE International Conference on Acoustics. 2017.
[13]Liang L, Zhang X, Renais S. On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition[C]// IEEE International Conference on Acoustics. 2016.
[14]Singhal S, Passricha V, Sharma P, et al. Multi-level region-of-interest CNNs for end to end speech recognition[J]. Journal of Ambient Intelligence and Humanized Computing, 2018.
[15]Wang X, Zhang P, Zhao Q, et al. Improved End-to-End Speech Recognition Using Adaptive Per-Dimensional Learning Rate Methods[J]. Ieice Transactions on Information & Systems, 2016, 99(10):2550-2553.
[16]Ito H, Hagiwara A, Ichiki M, et al. End-to-end neural network modeling for Japanese speech recognition[J]. Acoustical Society of America Journal, 2016, 140(4):3116-3116.
[17]Lu C. A Front-End Solution to VAD, SNR, and AGC for Speech Recognition[J]. 2009.