International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi :https://doi.org/10.32628/IJSRST229149

A Survey Paper on Speech Recognition System with Improved CLDNN Structure

Harshal Jagannath Jaware

Department of Computer Engineering (Data Science), Zeal College of Engineering and Research, Narhe, Pune, Savitribai Phule Pune University, Pune, Maharashtra, India

Article Info
Volume9, Issue 2
Page Number: 529-534

Publication Issue
March-April-2022

Article History
Accepted :03April2022
Published :20April2022

ABSTRACT

In the field of end-to-end speech recognition technology based on deep learning, CLDNN (Convolutional Long Short-Term Memory Fully Connected Deep Neural Network) is a commonly used model structure. The fully connected LSTM (Long Short Term Memory) model is used in the traditional CLDNN structure to process the timing information in the speech signal, which is prone to overfitting during the training process and affects the learning effect. Deeper models tend to perform better, but increasing the model depth by Simply stacking the network layers can cause gradient disappearance, gradient explosion, and "degeneration" problems. Aiming at the above phenomena and problems, this paper proposes an improved CLDNN structure. It combines the residual network and ConvLSTM to establish the residual ConvLSTM model, and replaces the fully connected LSTM model in the traditional CLDNN structure. The model structure solves the problems of the traditional CLDNN model, and can increase the model depth by stacking residual ConvLSTM blocks without gradient disappearance, gradient explosion and degeneration problems, which makes the speech recognition system perform better. The experimental results show that the model structure has a word error rate (WER) decrease of more than 8% in both Chinese and English speech recognition tasks compared to the traditional CLDNN structure.

I. INTRODUCTION

At present, relevant technology companies at home and abroad are constantly developing their own end-to-end speech recognition model.In recent years, thanks to the breakthrough of deep learning, automatic speech recognition technology is also in the stage of rapid development. The end-to-end speech recognition system based on deep learning has surpassed the traditional speech recognition system in popularity in academia, and began to gradually replace the traditional speech recognition system for practical production.Both of them use CLDNN and CTC to build speech recognition models and achieve excellent performance. DFCNN can see very long history and future information by accumulating a lot of these convolution pooling layer pairs, which ensures that DFCNN can express the long term correlation of voice excellently, and is more robust than RNN network structure. According to an article published by IBM researchers in ICASSP in 2016, using 3x3 convolution cores and multi-layer convolution followed by pooling layers, 14 layers Deep CNN models can be trained. Compared with the traditional CNN usage model, this model can bring about a relative 10.6% decline in WER on the Swatboard data set. The MSRA team proposed the residual network in 2015, which solved the degradation problem as the depth of the model deepened. Residual network has been applied to speech recognition models and proved to be effective. At the icassp conference in 2017, the Google research team presented an acoustic model structure combined with Network-in-Network (NiN), Batch Normalization (BN) and ConvLSTM. In the absence of a language model, the model achieves 10.5% WER in WSJ speech recognition tasks. CLDNN is a popular structure in the end-to-end speech recognition model because of its simple structure and excellent performance. But the common CLDNN model is not deep enough, and the extracted features are not rich enough, so the speech recognition model can not achieve the best effect. The fully connected long-term and short-term memory model (FC-LSTM) can not maintain the structural locality of speech feature space and is easy to over-fit. To solve these problems, Convolutional Long Short-Term Memory (ConvLSTM) is introduced to replace FC-LSTM in the CLDNN model, which improves the problem that the model can not maintain the locality of spatial structure and is easy to over-fit. In order to deepen the depth of the model without degradation, gradient disappearance and gradient explosion, Residual Network (ResNet) is also introduced in some papers. Based on the above improvements, an end-to-end speech recognition model based on CNN-ResConvLSTM-DNN and CTC structure is proposed by researchers. Compared with the CLDNN model, the WER of this model is 8.90% and 8.78% lower in Chinese and English speech recognition tasks.

II. METHODOLOGIES

a. General CLDNN Model

The general structure of the CLDNN network model is shown in Figure 1. The input layer is a frame-level acoustic feature related to the frequency domain. The frame-level feature is input to several layers of CNN for frequency convolution to reduce the frequency domain change.

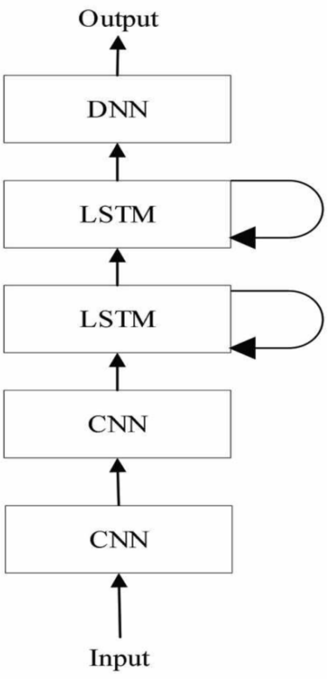


Fig. 1. General Structural Diagram of CLDNN

The output of the last layer of CNN is input to several layers of LSTM to provide context long-term memory. The output of the last layer of LSTM is input to the DNN layer. The purpose is to map the feature space to the output layer which is easier to classify.

b. Improved CLDNN Structure

1. Connectionist Temporal Classification

Traditional acoustic model training is based on frame-level labels with cross entropy criterion, which requires a tedious label alignment process. In order to achieve the goal of end-to-end training, the CTC objective function is used to automatically learn the alignment between network output and label sequence.

2. Convolutional LSTMNetwork(ConvLSTM)

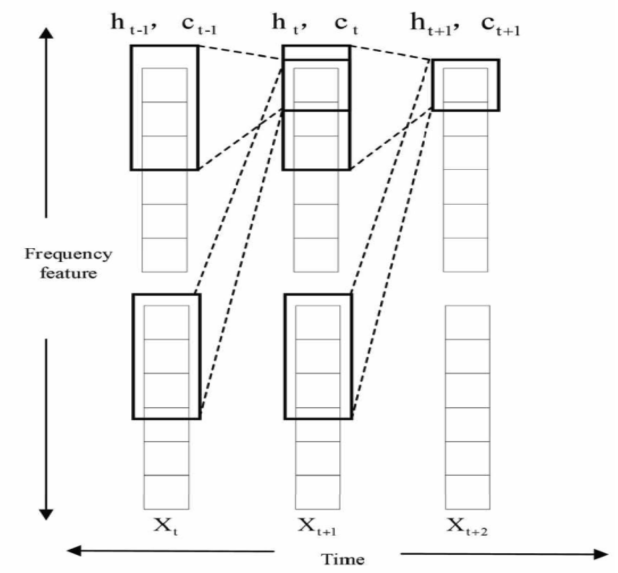


fig 2. The Internal Structure of ConvLSTM

The fully connected long-term memory model (FC-LSTM) has proven to be very effective in processing time correlation, but it does not maintain the structural locality of the features and is prone to overfitting. This paper introduces a convolutional long-term memory model, which is an extension of a fully connected long-term and short-term memory model. It has a convolution structure in both input state and state to state transitions. This structure is more capable than ordinary CNN. Time relationship, and it is less likely to overfit than fully connected LSTM

3. Residual Networks

When building a network, the deeper the network depth, the richer the feature hierarchy that can be extracted, so deeper networks can achieve higher levels of features and achieve better results in a given task. However, when constructing deep networks, gradient disappearance, gradient explosion, and degeneration problems are often encountered, resulting in deeper model training difficulties. This paper introduces the residual network structure to construct the deep network, and directly connects the shallow network and the deep network through the skip connection, so that the gradient can be better transmitted to the shallow layer to solve the above problems.

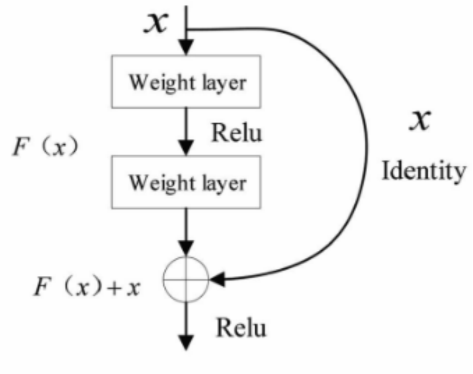


Fig 3. The Residual Block Structure

4. CNN-ResconvLSTM-DNN structure

In order to stack the multi-layer ConvLSTM to improve the performance of the model without gradient disappearance, gradient explosion and degeneration problems, this paper combines ConvLSTM and residual network structure.

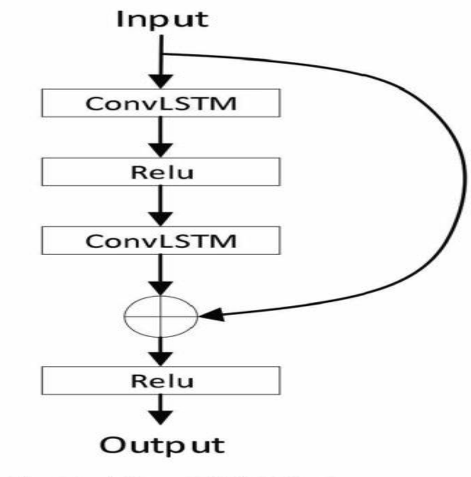


Fig 4. Residual ConvLSTM Block

Based on the above structure, this paper proposes an improvement on the traditional CLDNN structure. Aiming at the problem that the fully connected long-term memory model in the traditional CLDNN model can not maintain the structural locality of the feature space and is easy to over-fitting, the deep residual ConvLSTM network structure composed of multiple residual ConvLSTM blocks is used to replace the traditional CLDNN model. The multi-layer LSTM structure in the model gives the model a better representation of the temporal relationship in the processing of speech features and is less prone to overfitting. The improved CNN-ResconvLSTM-DNN model can build deeper models by superimposing more residual ConvLSTM without gradient disappearance, gradient explosion and degeneration problems, and can achieve better performance in speech recognition tasks. Its structure is as

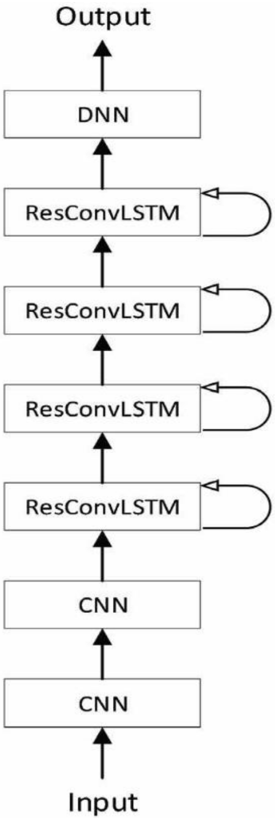


Fig 5. Cnn-Resconvlstm-Dnn Model Structure

III. RESULT AND ANALYSIS

A. Database

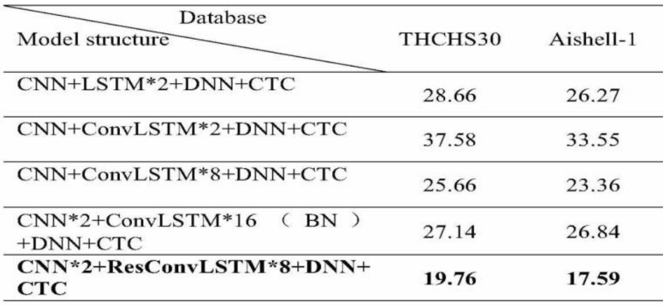
They evaluate the model on Chinese and English data sets. The Chinese task selects the THCHS30 corpus and the Aishell-1 corpus for experiments. The corpus collected byTHCHS30 includes a training set with a duration of 25 hours and 10,000 sentences, a development set with 2.14 hours duration, 893 sentences, and a test set with 6.15 hours duration and 2495 sentences. The person involved in the recording is a college student who speaks fluent Mandarin. The sampling frequency is 16 kHz and the sampling size is 16 bits. Aishell-1 collects 178 hours of corpus, covering 11 areas including smart home, driverless, industrial production, recording in a quiet environment, using three different devices: high-fidelity microphone, Android system phone and IOS The system phone has a sampling frequency of 16 kHz and a sampling size of 16 bits. The English task selects the TIMIT and Switchboard-1 corpus for experiments. The TIMIT corpus contains 6,300 given sentences from 630 people from eight major dialect regions of the United States, with a sampling rate of 16 kHz and a sample size of 16 bits. The official divides 70% of the corpus into training sets and 30% into test sets. The Switchboard-1 corpus collected 260 hours of recordings of 2,400 telephone conversations by 543 callers, with a sampling frequency of 8kHZ and a sampling size of 16 bits.

B. Method

The acoustic features use 40-dimensional MFCC. The number of ConvLSTM nodes in each residual block is 512, the initial learning rate is 0.001, the decay rate of learning rate is 50%, the batch size is 100, and a total o f 1000 training sessions.

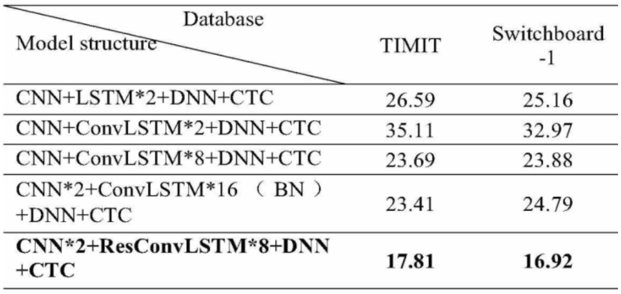
C. Result and Analysis

The Chinese test results are shown in Table 1, and the English test results are shown in Table 2



Database Model structure	THCHS30	Aishell-1
CNN+LSTM*2+DNN+CTC	28.66	26.27
CNN+ConvLSTM*2+DNN+CTC	37.58	33.55
CNN+ConvLSTM*8+DNN+CTC	25.66	23.36
CNN*2+ConvLSTM*16 （ BN ） +DNN+CTC	27.14	26.84
CNN*2+ResConvLSTM*8+DNN+ CTC	**19.76**	**17.59**

1. Performance of model in Chinese Recognition task (WER%)



Database Model structure	TIMIT	Switchboard -1
CNN+LSTM*2+DNN+CTC	26.59	25.16
CNN+ConvLSTM*2+DNN+CTC	35.11	32.97
CNN+ConvLSTM*8+DNN+CTC	23.69	23.88
CNN*2+ConvLSTM*16 （BN） +DNN+CTC	23.41	24.79
CNN*2+ResConvLSTM*8+DNN +CTC	**17.81**	**16.92**

2. **Performance of model in English Recognition task (WER%)**

Chinese Speech Recognition Tasks Use Words To Classify. As Can Be Seen From Table 1, In The Shallow Case, Convlstm Does Not Perform As Well As Fully Connected Lstm. As The Network Model Deepens, Convlstm Performs Better And Better When Convlstm Is Increased To About 16 Layers, The Regularization Layer Is Added To Overcome The Problem Of Gradient Disappearance And Gradient Explosion Generated By The Model. The Network Model Has A "Degraded" Phenomenon, And The Performance Is Not As Shallow As The Model. The Optimal Model Established In This Paper Is A Deep Network Model With Eight Residual Convlstm Blocks, Which Overcomes The Gradient Disappearance, Gradient Explosion And Degeneration Phenomenon Caused By Network Depth. It Has 8.9% And 8.68% of the WER Decline In Two Test Sets Compared With The Traditional Cldnn Model. English Speech Recognition Tasks Use Phonemes For Classification. It Can Be Seen From Table 2 That The Performance Results Of Each Model In The English Recognition Task Are Similar To Those In The Chinese Recognition Task. The Optimal Model Structure Designed In This Paper Has 8.78% And 8.24 % Of Wer Drops On The Two Test Sets Compared With The Traditional Cldnn Model.

IV. CONCLUSION

In view of the problem that the fully connected LSTM in the traditional CLDNN model is easy to over-fitting the model, and the problem of gradient disappearance, gradient explosion and degeneration occurs by simply superimposing the network layer to increase the depth of the model, the following research progresses:

1. The residual ConvLSTM model is used to replace the fully connected LSTM model in the traditional CLDNN model, and an improved CLDNN model structure is proposed.
2. Through the Chinese and English speech recognition experiments, the structure can effectively solve the problems existing in the traditional CLDNN, and can overcome the gradient disappearance, gradient explosion and degeneration problems caused by increasing the depth of the model, and also perform on the speech recognition task. Better than the traditional CLDNN model.

Due to the increase in depth, the training time required for this model is increased compared with the traditional CLDNN. The future research work will focus on further optimizing the model structure, shortening the training time as much as possible without reducing the performance of the model, and striving to achieve or approach the time required for traditional CLDNN training under the same task.

V. REFERENCES

[1]. Siniscalchi S M, Yu D, Deng L, et al. “Exploiting Deep Neural Networks for Speech Recognition”. Neurocomputing, 2013, 106(12): 148-157

[2]. Yang Y, Wang Y. “Speech recognition based on improved convolutional neural network algorithm”. Journal of Applied Acoustics, 2018, 37(06) : 1-7.

[3]. Li H, Jian S, Xu Z, et al. “Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network”. IEEE Transactions on Multimedia, 2017, 19(12):1-1

[4]. Xingjian S H I , Chen Z, Wang H, et al. “Convolutional LSTM network: A machine learning approach for precipitation nowcasting

Advances in neural information processing systems". 2015: 802-810.

[5]. Hannun A, Case C, Casper J, et al. "Deep speech: Scaling up end-to-end speech recognition International Conference on Machine Learning." 2014.

[6]. Graves A, Jaitly N. "Towards End-To-End Speech Recognition with Recurrent Neural Networks International Conference on Machine Learning". 2014.