MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Cell & Cellular Life Sciences Journal
ISSN: 2578-4811

Application of Hybrid CTC/2D-Attention End-to-End Model in Speech Recognition during the COVID-19 Pandemic

Bin Zhao[1]*, Mingzhe E[1] and Xia Jiang[2]

[1]School of Science, Hubei University of Technology, China

[2]Hospital, Hubei University of Technology, China

***Corresponding author:** Bin Zhao, School of Science, Hubei University of Technology, Wuhan, Hubei, China, Tel/Fax: +86 130 2851 7572; Email: zhaobin835@nwsuaf.edu.cn

Research Article

Volume 6 Issue 2

Received Date: August 18, 2021

Published Date: September 17, 2021

DOI: 10.23880/cclsj-16000163

Abstract

Recent research in the field of speech recognition has shown that end-to-end speech recognition frameworks have greater potential than traditional frameworks. Aiming at the problem of unstable decoding performance in end-to-end speech recognition, a hybrid end-to-end model of connectionist temporal classification (CTC) and multi-head attention is proposed. CTC criterion was introduced to constrain2D-attention, and then the implicit constraint of CTC on 2D-attention distribution was realized by adjusting the weight ratio of the loss functions of the two criteria. On the 178h Aishell open source dataset, 7.237% word error rate was achieved. Experimental results show that the proposed end-to-end model has a higher recognition rate than the general end-to-end model, and has a certain advance in solving the problem of mandarin recognition.

Keywords: Speech Recognition; 2-Dimensional Multi-Head Attention; Connectionist Temporal Classification; COVID-19

Introduction

Speech recognition technology is one of the important research directions in the field of artificial intelligence and other emerging technologies. Its main function is to convert a speech signal directly into a corresponding text. Yu Dong et al. proposed deep neural network and hidden Markov model, which has achieved better recognition effect than GMM-HMM system in continuous speech recognition task [1]. Then, Based on Recurrent Neural Networks (RNN) [2-5] and Convolutional Neural Networks (CNN) [6-11], deep learning algorithms are gradually coming into the mainstream in speech recognition tasks. And in the actual task they have achieved a very good effect. Recent studies have shown that end-to-end speech recognition frameworks have greater potential than traditional frameworks. The first is the Connectionist Temporal Classification (CTC) [12], which enables us to learn each sequence directly from the end-to-end model in this way. It is unnecessary to label the mapping relationship between input sequence and output sequence in the training data in advance so that the end-to-end model can achieve better results in the sequential learning tasks such as speech recognition. The second is the encode-decoder model based on the attention mechanism. Transformer [13] is a common model based on the attention mechanism. Currently, many researchers are trying to apply Transformer to the ASR field. Linhao Dong, et al. [14] introduced the Attention mechanism from both the time domain and frequency domain by applying 2D-attention, which converged with a small training cost and achieved a good effect. And Abdelrahman Mohamed [15] both used the characterization extracted from the convolutional network to replace the previous absolute position coding representation, thus making the feature length as close as possible to the target output length, thus saving calculation and alleviating the mismatch between the length of the feature sequence and the target sequence. Although the effect is not as good as the RNN model [16], the word error rate is the lowest in the method without language model. Shigeki Karita, et al. [17] made a complete comparison between RNN and Transformer in multiple languages, and the performance of Transformer has certain advantages in every task. Yang Wei, et al. [18] proposed that the hybrid architecture of CTC+attention has certain advancement in the task of Mandarin recognition

with accent. In this paper, a hybrid end-to-end architecture model combining Transformer model and CTC is proposed. By adopting joint training and joint decoding, 2D-Attention mechanism is introduced from the perspectives of time domain and frequency domain, and the training process of Aishell dataset is studied in the shallow encoder-decoder network.

Hybrid CTC/Transformer Model

The overall structure of the hybrid CTC/Transformer model is shown in Figure 1. In the hybrid architecture, chained chronology and multi-head Attention are used in the process of training and grading, and CTC is used to restrain Attention and further improve the recognition rate.

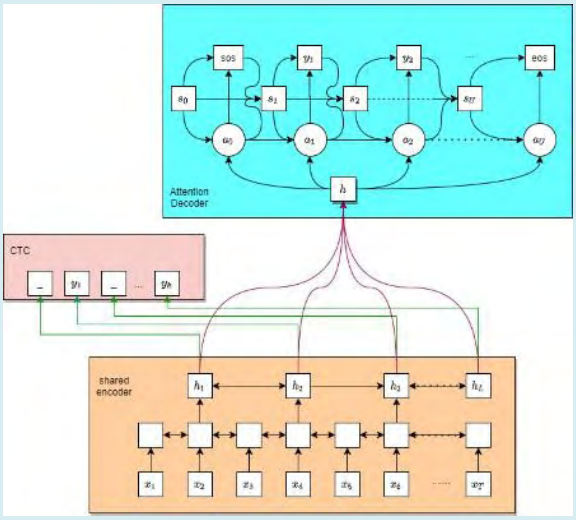

Figure 1 An encoder-decoder architecture based on Transformer and CTC.

In end-to-end speech recognition task, the goal is through a network, the input $X=(X_1,....X_T)$ calculate all output tags sequence $Y=(Y_1,....,Y_M)$ corresponding probability, usually $M \le T$, $Y_m \in L$, L is a finite character set, the final output is one of the biggest probability tags sequence, namely

$$y^* = \frac{\text{argmax } P(y/x)}{y} \quad (1)$$

Connectionist Temporal Classification

Connectionist Temporal Classification structure as shown in Figure 2, in the training, can produce middle sequence $\pi = (\pi_1,...,\pi_0)$, in sequence π allow duplicate labels, and introduce a blank label *blank*: <–> have the effect of separation, namely $\pi_i \in L \cup \{blank\}$. For example y = (wo,ai,ni,Zhong,guo), π = (-, wo, - , -,ai,ai, - ,ni,ni, - ,zhong,guo, –) , y' = (- ,wo, - ,ai, - ,ni, - ,zhong, - ,guo, -), this is equivalent to construct a many-to-one mapping $B:L' \to L^{\le T}$, The $L^{\le T}$ is a possible π output set in the middle of the sequence, and then get the probability of final output tag:

$$P(y \mid x) = \sum_{\pi \in B^{-1}(y')} P(\pi \mid x) \quad (2)$$

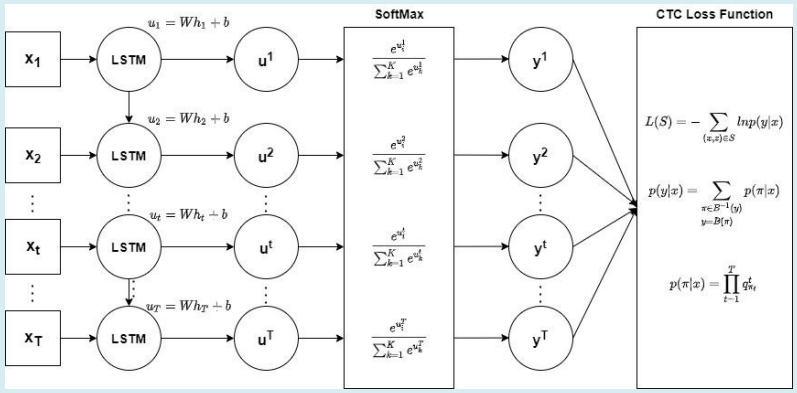

Figure 2: Connectionist Temporal Classification.

Where, S represents a mapping between the input sequence and the corresponding output label; $q^t_{\pi_t}$ represents label π_t at time t corresponding probability. All tags sequence in calculation through all the time, because of the need to N^T iteration, N said tag number, the total amount of calculation is too big. HMM algorithm can be used for reference here to improve the calculation speed:

$$P_{ctc}(\mathrm{y}/\ \mathrm{x}) = \sum_{t=1}^{T} \sum_{u=1}^{|Y'|} \frac{\alpha_t(u)\beta_t(u)}{q^t_{\pi_t}} \quad (3)$$

Among them, the $\alpha_t(u)$、$\beta_t(u)$ respectively are forward probability and posterior probability of the -th label at the moment to. Finally from the intermediate sequence to the output sequence, CTC will first recognize the repeated characters between the delimiters and delete them, and then remove the delimiter <–>.

Transformer Model for 2D-Attention

Transformer model is used in this paper, which adopts a multi-layer encoder-decoder model based on multi-head attention. Each layer in the encoder should be composed of a 2-dimensional multi-head attention layer, a fully connected network, and layer normalization and residual connection. Each layer in the decoder is composed of a 2-dimensional multi-head attention layer that screens the information before the current moment, an attention layer that calculates the input of the encoder, a three-layer network that is fully connected, and a layer normalization and residual connection. The multi-head attention mechanism first initializes the three weight matrices Q.K.V by means of linear transformation of the input sequence:

$$\mathrm{Q} = \mathrm{W_Q X};\ \mathrm{K} = \mathrm{W_K X};\ \mathrm{V} = \mathrm{W_V X} \quad (4)$$

Then the similarity between the matrix and K is calculated by dot product:

$$f(\mathrm{Q},\mathrm{K}) = \frac{QK^T}{\sqrt{d_k}} \quad (5)$$

In the process of decoder, requiring only calculated before the current time and time characteristics, the similarity between the information on subsequent moment for shielding, usually under the introduction of a triangle total of 0 and upper triangular total negative infinite matrix, then, the matrix calculated by Equation (5) is transformed into a lower triangular matrix by replacing the negative infinity in the final result with 0. Finally, Softmax function is used to normalize the output results, and weighted average calculation is carried out according to the distribution mechanism of attention:

$$Attention\ (Q,K,V) = softmax\ (f\ (Q,K))\ V \quad (6)$$

Multi-head attention mechanism actually is to multiple independent attention together, as an integrated effect, on the one hand, can learn more information from various angles, on the one hand, can prevent the fitting, according to the calculation of long attention, if the above results when the quotas for time calculation, then the final result stitching together, converted into a linear output:

$$multiHead\ (Q,K,V) = concat\ (head_{1,\ldots\ldots,} head_n)W^0 \quad (7)$$

2D - Attention: The Attention structure in Transformer only models the position correlation in the time domain. However, human beings rely on both time domain and frequency domain changes when listening to speech, so the 2D-Attention structure is applied here, as shown in Figure 3, that is, the position correlation in both time domain and frequency domain is modeled. It helps to enhance the invariance of the model in time domain and frequency domain.

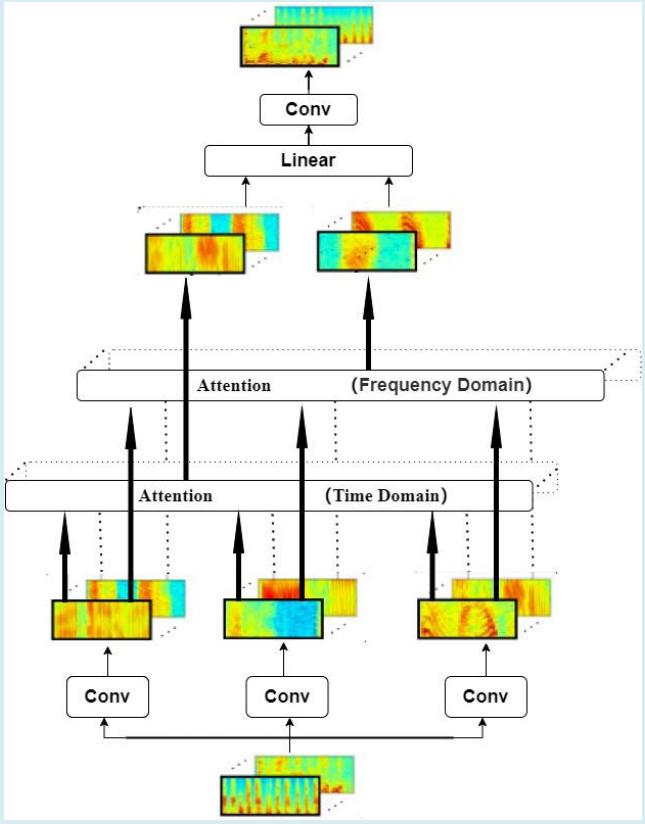

Figure 3: The structure of 2D-Attention.

Its calculation formula is as follows:

$$2D - Attention(I) = W^0 * concat(channel_1^f, \ldots\ldots, channel_c^f, channel_1^t, \ldots\ldots, channel_c^t) \quad (8)$$

Where

$$channel_1^f = attnetion((W_i^Q * I)^T, (W_i^K * I)^T, (W_i^V * I)^T)$$

$$channel_1^t = attnetion((W_i^Q * I),(W_i^K * I),(W_i^V * I))$$

After calculating the 2-dimensional multi-head attention mechanism, a feed-forward neural network is required at each layer, including a fully connected layer and a linear layer. The activation function of the fully connected layer is ReLU:

$$FFN(X) = max(0, xW_1 + b_1)W_2 + b_2 \quad (9)$$

In order to prevent the gradient from disappearing, the residual connection mechanism should be introduced to transfer the input from the bottom layer directly to the upper layer without passing through the network, so as to slow down the loss of information and improve the training stability:

$$x + SubBlock(LayerNorm)(x)) \quad (10)$$

To sum up, the Transformer model with 2D-attention is shown in Figure 4.

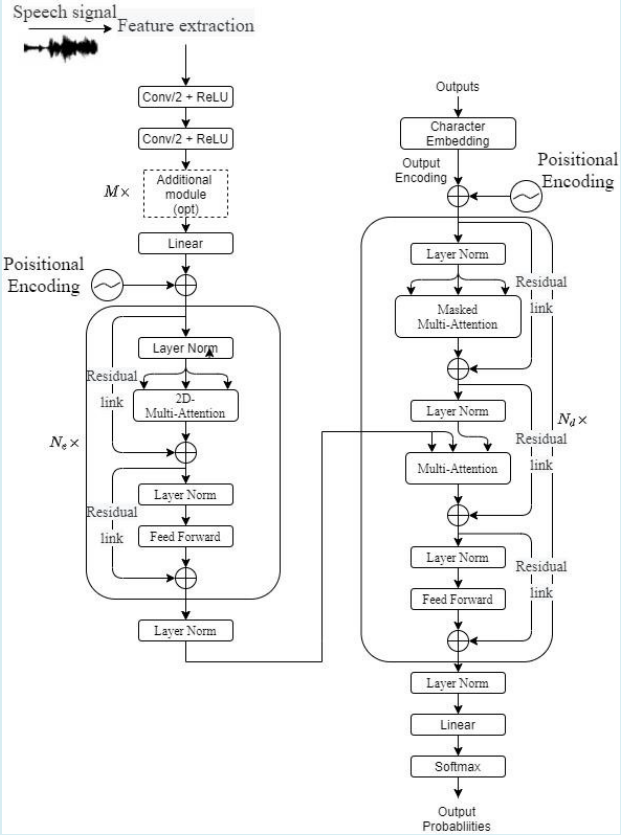

Figure 4: Overall network architecture of Transformer model with 2D-Attention. The loss function is constructed according to the principle of maximum likelihood estimation:

The speech features are first convolved by two operations, which on the one hand can improve the model's ability to learn time-domain information. On the other hand, the time dimension of the feature can be reduced to close to the quotas-length of the target output, which can save calculation and alleviate the mismatch between the length of the feature sequence and the target sequence.

$$L_{att} = -logp(y_1, y_2, ..., y_{T'} | x_1, x_2, ..., x_T)$$

$$\sum_{t'=1}^{T'} P(y_{t'} | y_1, y_2, ..., y_{t'-1}, Z \quad (11)$$

The final loss function consists of a linear combination of CTC and Transformer's losses:

$$L = \mu L_{CTC} + Y^{L_{att}} \quad (12)$$

FBANK Feature Extraction

The process of FBANK feature extraction is shown in Figure 5. In all experiments, the sampling frequency of 1.6KHz and the 40-dimensional FBANK feature vector are adopted for audio data, 25ms for each frame and 10ms for frame shift.

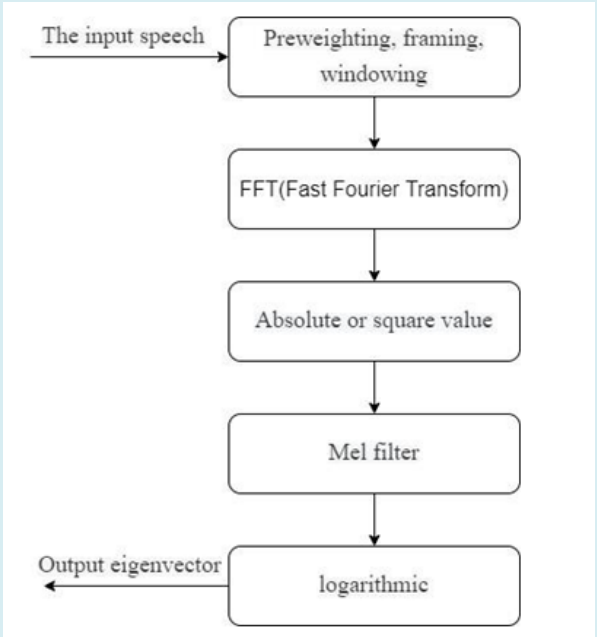

Figure 5: Flowchart of FBANK feature extraction.

Speech Feature Enhancement

In computer vision, there is a feature enhancement method called "Cutout" [19,20]. Inspired by this method, this paper adopts time and frequency shielding mechanism to

screen a continuous time step and MEL frequency channel respectively. Through this feature enhancement method, the purpose of overfitting can be avoided. The specific methods are as follows:

- **Frequency shielding:** Shielding f consecutive MEL frequency channels: $[f_0, f_0 + f)$, replace them with 0. Where, is from zero to custom frequency shielding parameters randomly chosen from a uniform distribution, and f_0 is selected from [0,v-f) randomly, v is the number of MEL frequency channel.

- **Time shielding:** Time step $[t_{0'0} + t)$ for shielding, use 0 for replacement, including from zero to a custom time block parameters randomly chosen from a uniform distribution, t_0 from $[0,r - t)$ randomly selected. The speech characteristics of the original spectra and after time and frequency shield the spectrogram characteristic of language contrast as shown in Figure 6, to achieve the purpose of to strengthen characteristics.

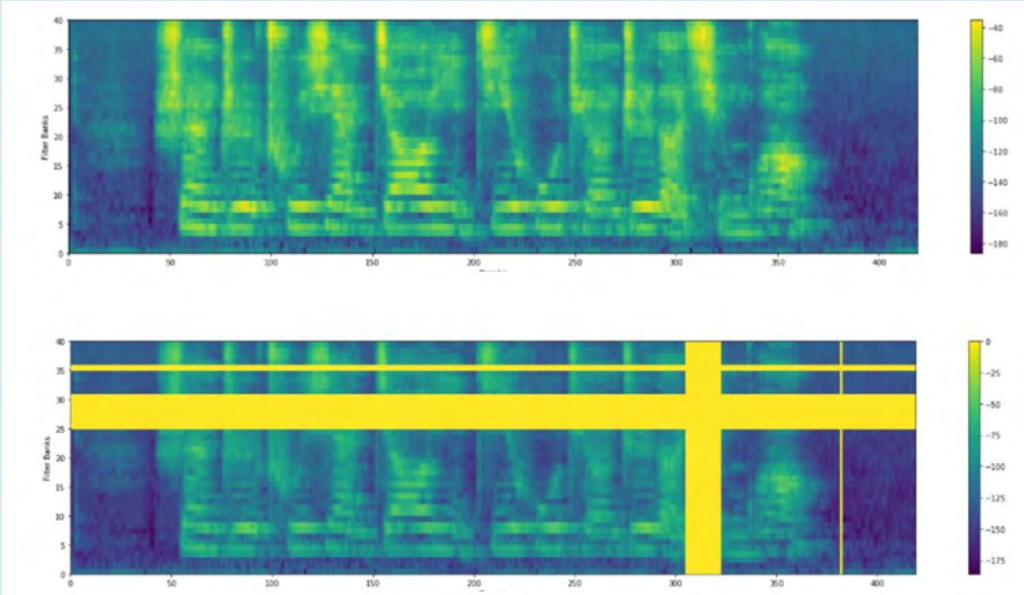

Figure 6: Feature enhancement contrast chart.

Label Smoothing

The learning direction of neural network is usually to maximize the gap between correct labels and wrong labels. However, when the training data is relatively small, it is difficult for the network to accurately represent all the sample characteristics, which will lead to overfitting. Label smoothing solves this problem by means of regularization. By introducing a noise mechanism, it alleviates the problem that the weight proportion of the real sample label category is too large in the calculation of loss function, and then plays a role in inhibiting overfitting. The true probability distribution after adding label smoothing becomes:

$$p_i = \{\frac{1,\ if\,(i=y)}{0,\ if\,(i\neq y)} \Rightarrow p_i\{\frac{(1-\varepsilon),\ if\,(i=y)}{\frac{\varepsilon}{k-1},\ if\,(i\neq y)}$$

Where K represents the total number of categories of multiple classifications, and is a small hyperparameter.

Experiment

Experimental Model

The end-to-end model adopted in this paper is a hybrid model based on Linked Temporalism and Transformer based on 2-dimensional multi-head attention. Compare the end-to-end model based on RNN-T and the model based on multiple heads of attention. The experiment was carried out under the Pytorch framework, the GPU RTX 3080.

Data

This article uses Hill Shell's open source Aishell dataset, which contains about 178 hours of open source data. The dataset contains almost 400 recorded voices from people with different accents from different regions. Recording was done in a relatively quiet indoor environment using three different devices: a high-fidelity microphone (44.1kHz, 16-bit); IOS mobile devices (16kHz, 16-bit); Android mobile

device (16kHz, 16-bit) to record, and then by sampling down to 16kHz.

Network Structure and Parameters

The network in this paper uses four layers of multi-head attention, and the input attention dimension of each layer is 256, the input feature dimension of the forward full connection layer is 256, and the hidden feature dimension is 2048. The combined training parameter λ is 0.1, the rate of random loss of activated cells is 0.1, and the label smoothing is 0.1. The epoch times are 200.

Evaluation Index

In the evaluation of experimental results, word error rate (WER) was used as the evaluation index. Word error rate is identified primarily for the purpose of make can make between words and real words sequences of the same, the need for specific words, insert, substitute, or delete these insertion (I), substitution (S) or deletion (D) of the total number of words, divided by the real word sequence of all the percentage of the total number of words. Namely

$$WER = 100\,\mathrm{X}\,\frac{I+S+D}{N}\%$$

Experimental Results and Analysis

Table 1 shows the comparison between the 2D-attention model without CTC and the 2D-attention model with CTC. Compared with the ordinary model, the performance of the model with CTC is improved by 6.52%-10.98%, and the word error rate of the end-to-end model with RNN-T is reduced by 4.26%. Performance improved by 37.07%.

Model	Test-WER/%
RNN-T	11.5
4Enc+3Dec	9.32
4Enc+4Dec	9.165
4Enc+4Dec+0.1CTC	8.567
6Enc+3Dec	8.13
6Enc+3Dec+0.1CTC	**7.237**

Table 1: Comparison of model word error rate.

Figure 7 shows the comparison of the loss functions of the two models (2D-attention model without CTC and 2D-attention model with CTC). Compared with the ordinary model, the loss of the constrained model with CTC can reach a smaller value.

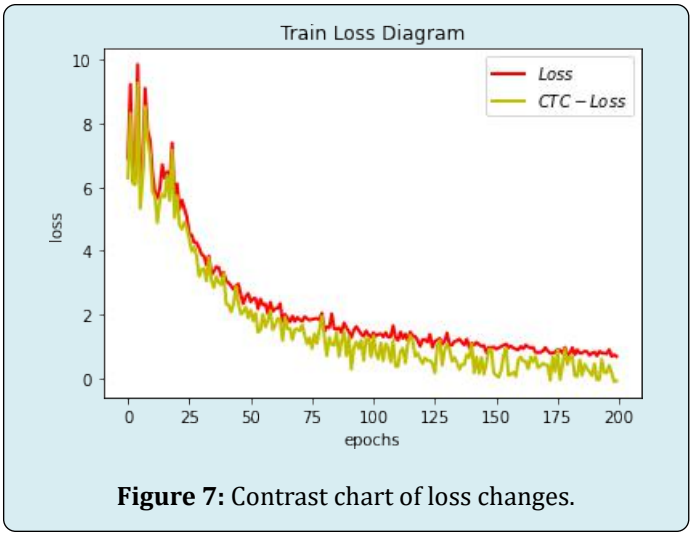

Figure 7: Contrast chart of loss changes.

Conclusion

In this paper, we propose a hybrid architecture model of Transformer using CTC and dimensional Multi-head Attention to apply to Mandarin speech recognition. Compared with the traditional speech recognition model, it does not need to separate the acoustic model and the language model to train, only needs to train a single model, from the time domain and frequency domain two perspectives to learn the speech information, can achieve advanced model recognition rate. It is found that compared with the end-to-end model of RNN-T, the performance of Transformer model is better, and the increase in the depth of the encoder can better learn the information contained in the speech, which can significantly improve the performance of the model for Mandarin recognition, while the increase in the depth of the decoder has little effect on the overall performance of speech recognition. At the same time, by introducing a link of sequence alignment is improved, the model makes the model to achieve the best effect, but on some professional vocabulary is not accurate, the subsequent research by increasing solution of language model, at the same time, in view of the very deep network training speed slow problem, to improve and upgrade.

Conflict of Interest

We have no conflict of interests to disclose and the manuscript has been read and approved by all named authors.

Acknowledgment

This work was supported by the Philosophical and Social Sciences Research Project of Hubei Education Department (19Y049), and the Staring Research Foundation for the Ph.D.

of Hubei University of Technology (BSQD 2019054), Hubei Province, China.

References

1. Yu D, Deng L, YU Kai (2016) Analytical Deep Learning: Practice of Speech Recognition , Beijing: Publishing House of Electronics Industry.

2. Dahl GE, Yu D, Deng L, Acero A (2012) Context-dependent pre-train deep neural networks for large vocabulary speech recognition. IEEE Transactions on 20(1): 30-42.

3. Hinton G, Deng L, Yu D, Dahl G, Mohamed A, et al. (2012) Deep Neural Networks for Acoustic modeling in Speech Recognition: The Shar Views of Four Research Groups. Signal Processing Magazine, IEEE 29(6): 82-97.

4. Hochreiter S, Schmidhuber J (1997) Long Short-term Memory. Neural Computation 9(8): 1735-1780.

5. Zhang Y, Chen GG, Yu D, Yao K, Kundanpur S, et al. (2016) Highway Long Short-term Memory RNNS for Distant Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, March 20-25, Shanghai, China. Piscataway.

6. Lecun Y, Bengio Y (1995) Convolutional Networks for Images, Speech and Time-series, Cambridge: MIT Press.

7. Abdel-Hamid O, Moham AR, Jiang H, Penn G (2012) Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM model for Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, pp: 4277-4280.

8. Abdel-Hamid O, Moham AR, Jiang H, Deng L, Penn G, et al. (2014) Convolutional Neural Networks for Speech Recognition. IEEE/ACM Transactions on Audio Speech & Language Processing 22(10): 1533-1545.

9. Abdel-Hamid O, Deng L, Yu D (2013) Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition. 25-29 August, Interspeech 58(4): 1173-1175.

10. Sainath TN, Moham AR, Kingsbury B, Ramabhadran (2015) Deep Convolutional Neural Networks for LVCSR. IEEE International Conference on Acoustics, Speech and Signal Processing, pp: 8614-8618.

11. Sainath TN, Vinyals O, Senior A, Sak H (2015) Convolutional, Long Short-term Memory, Fully Connect Deep Neural Networks. IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE Press, pp: 4580-4584.

12. Graves A, Fernández S, Gomez F, Schmidhuber JA (2006) Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Proceedings of the 23rd international conference on Machine learning, pp: 369-376.

13. Vaswani A, Shazeer N, Pamar N, Uszkoreit J, Jones L, et al. (2017) Attention is All You Need. Advances in Neural Information Processing Systems, pp: 6000-6010.

14. Dong L, Xu S, Xu B (2018) Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

15. Abdelrahman M, Dmytro O, Luke Z (2019) Transformers with Convolutional Context for ASR.

16. Kyu J H, Akshay C, Jungsuk K, Ian L (2017) The Capio 2017 Conversational Speech Recognition System.

17. Karita S, Chen N, Hayashi T, Hori T, Inaguma H, et al. (2019) A Comparative Study on Transformer vs RNN in Speech Applications. IEEE Automatic Speech Recognition and Understanding Workshop.

18. Yang W, Hu Y (2021) End-to-end Accent Putonghua Recognition Based on Hybrid CTC/ Attention Framework. Application Research of Computers 38(3): 755-759.

19. Pham NQ, Nguyen TS, Niehues J, Muller M, Stuker S, et al. (2019) Very Deep Self-attention Networks for End-to-end Speech Recognition.

20. Ekin D C, Barret Z, Dandelion M, Vijay V, Le QC (2018) Autoaugment: Learning Augmentation Policies from Data.