Advances in Intelligent Systems Research, volume 147
International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Continuous Speech Recognition Model Based on CTC Technology

Yumeng Wang [1, a)], Jianmin Zhao[2]

[1]*School of Computer and Information Technology, Northeast Petroleum University, Daqing 163000, China;*
[2]*School of Computer and Information Technology, Northeast Petroleum University, Daqing 163000, China.*

[a)] Corresponding author email: 1183650351@qq.com

Abstract. In end to end speech recognition, the linguistic knowledge such as pronunciation 1exicon is not essential. And therefore, the performance of the ASR systems based on CTC is weaker than that of the baseline, aiming at this problem, a strategy combining the existing linguistic knowledge and the acoustic modeling based on CTC is proposed and the tri—phone is taken as the basic units in acoustic modeling, Thus the sparse problem of the modeling unit is effectively solved and the discrimination and robustness of the CTC model are improved substantially.

Key words: Connectionist Temporical Classification; Speech Recognition; End to End.

INTRODUCTION

Faced with the age of big data, traditional machine learning algorithms have been unable to cope with massive raw speech data processing. Deep learning shows strong modeling ability in the field of pattern recognition and has become a hot research direction in speech recognition [1]. The deep belief neural network (DBNN) proposed by Hinton is a multilayer and densely connected neural network model. The advantage of DBNN is that by increasing the number of layers and nodes of the neural network, the ability of abstract generalization and modeling of the neural network to data is improved, but at the same time, DBNN also has some shortcomings. For the current speech frame processing, the splicing frame is generally adopted, which destroys the phase between the speech sequences. Customs. Graves proposed the training algorithm of connection time classification (CTC), which further improved the accuracy of phoneme recognition.

At present, the end-to-end system is usually composed of LSTM neural network and CTC loss function. Under the restriction of CTC algorithm [2], LSTM actively records the correspondence of acoustic features and related sequences during the training process. It is different from a series of assumptions based on the traditional HMM speech recognition. In the light of the characteristics of different languages [3], the end to end system is directly connected to the word and letter modeling on the basis of the traditional phonetics. Identification process. However, the complete abandonment of linguistic knowledge makes the end to end system have a certain gap between recognition rate and speech pattern recognition based on deep learning. Therefore, combining the advantages of CTC technology and neural network, a continuous speech recognition model is proposed, and the accuracy is further verified through experiments.

MIXED NEURAL NETWORK OF DBNN-BLSTM

Long Short-Term Memory.

Long Short-Term Memory (LSTM) neural network is a special case in the recurrent neural network (RNN). The basic idea of its nature to maintain the backpropagation error in a certain range of. LSTM is to use different types of doors to control the information flow of the neural network. LSTM neurons to preserve the information in the same way[4]. It can be stored for a long time. Through a variety of gate structures, LSTM can handle the information flow autonomously. A complete LSTM memory unit can be described by the following recursive formula:

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
c_t &= f_t.c_{t-1} + i\tanh(W_{xx}x_t + W_{hc}h_{t-i} + b_c) \\
h_t &= o_t\tanh(c_t)
\end{aligned}
$$

Among them: i, f, c, o are input gates, oblivion gates, neuron activation, output gates; W represents the weight of different gates; b is the corresponding bias matrix[5].

Bi-Directional Long Short-Term Memory.

Because of its structural characteristics, RNN can only handle current voice information, and cannot make use of historical information. However, in the process of speech recognition, the past historical context is also very important for the recognition accuracy. In order to solve this problem, the two-way BRNN uses two independent hidden layers to handle the data in the two directions, then feed forward to the same output layer, and the feedback path can be described by the following formula:

$$
\begin{aligned}
\overrightarrow{h}_t &= f(W_{\underset{xh}{\rightarrow}}x_t + W_{\underset{hh}{\rightarrow}}\overrightarrow{h}_{t-1} + b_{\underset{h}{\rightarrow}}) \\
\overleftarrow{h}_t &= f(W_{\underset{xh}{\leftarrow}}x_t + W_{\underset{hh}{\leftarrow}}\overleftarrow{h}_{t+1} + b_{\underset{h}{\leftarrow}}) \\
y_t &= W_{\underset{hy}{\rightarrow}}\overrightarrow{h}_t + W_{\underset{hy}{\leftarrow}}\overleftarrow{h}_t + b_y
\end{aligned}
$$

The flow direction of the two hidden layers is opposite when the input sequence information is forward. When all input sequences are completed, the output layer is updated. At the same time, the reverse output layer transfers the feedback information in the two opposite directions [6].

SPEECH RECOGNITION MODEL BASED ON CONNECTIONIST TEMPORAL CLASSIFICATION (CTC)

The speech recognition process is shown in Fig 1. First, the input training speech signal is preprocessed and feature extracted through neural network. At the same time, the acoustic model is established through training. The relationship between words or sentences is learned through language training, a language model is established, and the language model is used to estimate the input sequence of the test speech signal [7]. The possibility of the column is finally done by decoding the optimal results after the matching of the language model and the preprocessed and feature extracted test signals.

In Figure 1, the acoustic model is a BLSTM trained by the CTC algorithm, using the traditional 25ms frame length and 10ms frame shift to extract features and decoding it using the Vitby algorithm. First, the syntax-based language model is used to design, and it is successfully used to identify several thousands of concepts in the syntax state. The domain constraint task, the word recognition rate exceeds the 90%. Preprocessing Subsystem to achieve the primary level standard detection and speech detection by independent training [8]. The standardization is carried out in batch. Then the speech detection is carried out by the training mode similar to the acoustic model. Finally, the speech, nonsound and silent segments are identified. Test, use a new language model to complete the process of LVCSR. This language model combines three different language models, words, words, and grammar. Finally, the machine learning method is used to insert the model linearly and determine its interpolation weight, so as to minimize the complexity of the validation set. A method to find the best weight of the interpolation process by a method that minimizes the complexity of the final model on the development set is found by a method of minimizing the complexity of the final model on the development set [9]. The 40K word dictionary complexity is 376, and then they are used to annotate and eliminate the core training text and test sets. After using the new data retraining model, the complexity of the core test set is 246.

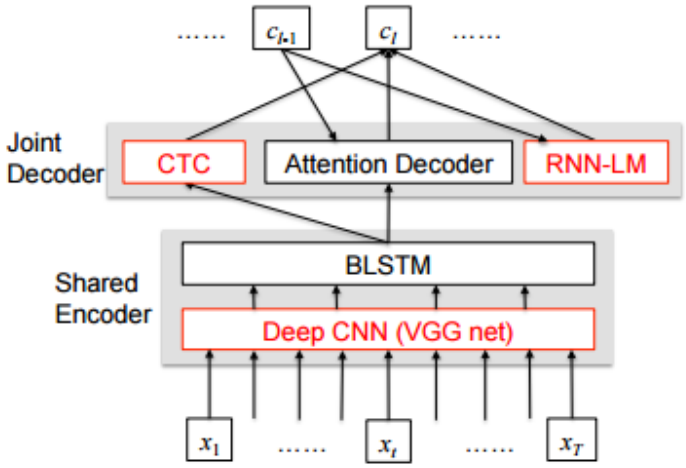

FIG. 1. The process of Speech recognition

EXPERIMENT AND RESULT ANALYSIS

The data set used in the experiment is the TIMIT data set of the English standard continuous speech recognition library [10]. All the data sets are divided into two sets: training set and test set. 90% of them are used for training and 10% for testing. Using MFCC to normalize speech features and use 36 target class tags. (36 phonemes). The DBNN is set to 7 layers, each layer of 1024 neurons, the DBNN training 75 cycles, the learning rate is set to 0.005, the momentum is set to 0.9. BLSTM-CTC initialization weights between 0.1 and 0.1 uniform random changes, its structure of 39 Input (speech feature) and about 113000 neurons [11].

Acoustic model	Acoustic database/h	Language template complexity	Recognition rate of words/%
BLSTM-CTC	120	246	65
DBNN-BLSTM	12	246	62
BLSTM-CTC	12	246	57

TABLE 1. Results of experiments

As can be seen from Table 1, the highest accuracy of word recognition is achieved by the BLSTM-CTC acoustic model trained on the corpus (120h), up to 65%. and BLSTM-CTC is only trained on the 12h corpus, and the rate of word recognition declines 8%. DBNN BLSTM in 12h The training results on the data base are 62% [12], and the recognition rate of BLSTM-CTC is 5% higher than that of the same time. Thus, the BLSTM-CTC is more capable of modeling the complex data than DBNN-BLSTM and is more advantageous in the continuous speech recognition of large vocabulary.

SUMMARY

The performance of the BLSTM-CTC model is obviously improved compared with the BLSTM neural network, but the work efficiency will be reduced because of the long training time of the large training set. The research on the speech recognition based on the CTC speech recognition model is still faced with many problems, such as the division of the acoustic model and the RUU. In the future, more research will be done on how to enhance the robustness of acoustic models [13].

REFERENCES

1. Zhou Shichao, Zhang Huyin, Yang Bing. Speech service text classification based on deep belief network. Computer engineering and applications, 2016 (21): 157 - 161.
2. Deng Kan, Ou Zhi Jian. Research on adaptive method for speech recognition in deep neural networks. Computer application research, 2016, 33 (7): 1966 - 1970.
3. Wang Xiaohua, Torre, Zhang Chao, et al. Speech feature extraction algorithm based on Bark wavelet packet transform based on Fisher ratio. Journal of Xi'an Polytechnic University, 2016, 30 (4): 452 - 457.
4. Zuo Lingyun, Zhang Qingqing, Li TA, et al. Reassessment method based on LSTM DNN language model in telephone conversation speech recognition. Journal of Chongqing University of Posts and Telecommunications (NATURAL SCIENCE EDITION), 2016, 28 (2): 180 - 186193.
5. Wang Bo. Approximation order of limit learning machine for convex increment. Journal of Xi'an Polytechnic University, 2015, 29 (6): 756 - 760.
6. single Yu Xiang, Deng Yan, Liu Jia. A new type of joint language recognition algorithm for large vocabulary continuous speech recognition. Journal of automation, 2012, 38 (3): 366 - 374.
7. Wang Shanhai, Jing Xin Yu, Yang Haiyan. Research on speech recognition of isolated words based on deep learning neural networks. Computer application research, 2015, 32 (8): 2289 - 2291.
8. Zhang Yunan, Liu Fuyong. An improved variable step size adaptive bat algorithm and its application. Journal of Guangxi University for Nationalities (NATURAL SCIENCE EDITION), 2013, 19 (2): 51 - 54.
9. Sheng Meng long, he Hsing, Wang Huimin, et al. An improved adaptive mutation bat algorithm. Computer technology and development, 2014, 24 (10): 131 - 134.
10. Tang Jianxin, Zhao Fuqing, Wang Xin, et al. Improved bats optimization algorithm based on the mechanism of velocity weighting perturbation. Journal of Lanzhou University of Technology, 2016, 42 (1): 104 - 108.
11. CHEN Si, he Hsing, Yang Xinshe, et al. Bats optimization algorithm based on time-varying inertia weight learning mechanism. Journal of basic science of Textile Universities, 2016, 29 (4): Five to five sixty.
12. Li Zhiyong, Ma Liang, Zhang Huizhen. Convergence analysis of the bat algorithm. The practice and understanding of mathematics, 2013, 43 (12): 182 - 190.
13. Sheng Meng long, he Xing, Ding Wenjing. The global convergence analysis of Ding Wenjing's bat algorithm. Journal of basic science of Textile Universities, 2013, 26 (4): 543 - 547.