The Implementation of Chinese Acoustic Model Efficiency Optimization Based on Convolutional Neural Network

Hao Zeng¹, Yinyin Deng², Can Huang³

¹ Chongqing University of Posts and Telecommunications, Chongqing, 400000, China
² Chongqing University of Posts and Telecommunications, Chongqing, 400000, China
³ Corresponding author’s e-mail: yin_zi_66@sina.com

Abstract. In the speech recognition system, the traditional convolution method of convolutional neural network (CNN) model has the problem of low efficiency, especially when processing a large number of speech feature map data. In this paper, a new convolution method for the CNN model has been proposed, which makes the efficiency of CNN model about five times faster than the traditional method. And at the same time, the recognition accuracy of the new method is comparable with traditional method.

1. Introduction
In the 21st century, with the rapid development of information technology, the role of human-computer interaction is more prominent. As the key technology of human-computer interaction, automatic speech recognition plays an increasingly important role. At present, the speech recognition system is still based on the statistical mode, and the whole recognition process can be roughly divided into several parts: front-end signal processing and speech feature extraction, acoustic model and language model, decoding search, and the like. As the underlying model of the speech recognition system, the acoustic model plays the most important role. The purpose of establishing the acoustic model is to provide a reliable method to calculate the distance between the feature vector sequence of the speech and the pronunciation template. At the same time, most of the computational overhead in speech recognition systems comes from acoustic model. The acoustic model largely determines the performance of the system, so choosing the appropriate acoustic model is very important for the speech recognition system.

At present, the Convolution Neural Networks (CNN) acoustic model is a mainstream model, it can achieve better recognition results, however, due to the large number of CNN model layers, the complexity of the model is too large. If the CNN model is directly used in the speech recognition system, it has rather low efficiency. Therefore, improving the efficiency of the model has become one of the main research directions in the speech recognition industry. This paper has done relevant research on this difficulty and achieved good results. The method in this paper has been deployed and used on related systems.

2. Introduction to acoustic model of convolutional neural networks
As a deep learning network model, CNN is better at capturing the non-deformation of speech features to better describe and learn speech signals. Using CNN acoustic model can greatly improve the accuracy of speech recognition. The basic network structure of CNN mainly includes input layer, convolution layer, pooling layer (also called down-sampling layer), fully connected layer and output
The convolutional layer and the pooling layer are alternately arranged, and in practical applications the total number of these two layers is from a dozen to dozens. The network structure of the CNN model is shown in Figure 1.

![CNN model structure](image)

### Figure 1. CNN model structure

Multiple sets of convolution kernels in the convolutional layer convolute the input feature map separately, we assume that the feature map $M$ has a size of $i \times i$, the convolution kernel $C$ has a size of $k \times k$, the extended edge is $p$, the convolution step is $s$, the offset variable is $b_0$, and the excitation function $f(x)$ is sigmoid function. Then the output feature map size is $O \times O$. The solution process of $O$ is as follows:

$$O = \left\lfloor \frac{i + 2p - k}{s} \right\rfloor + 1$$

[.] means rounding down, then the output of convolution calculation is as follows:

$$F_{ij} = f\left( \sum_{i=1}^{n} \sum_{j=1}^{m} (M_{ij} C_{ij}) + b_0 \right)$$

$F_{ij}$ is the individual elements of the output feature map. In addition to the sigmoid function, the commonly used excitation functions of the convolutional layer are tanh functions, ReLU functions, etc. In the convolution operation, convolution calculation is performed on different receptive fields by using the same convolution kernel. In this process, all the receptive fields are shared convolution kernel parameters. This operation of weight sharing greatly reduces the number of parameters during model training. If you want to further reduce the parameters, you can add a pooling layer after the convolutional layer for downsampling to make fewer parameters; the commonly used pooling methods mainly have maximum pooling and averaging pooling. When processing a large number of feature data, CNN acoustic model has better recognition accuracy than other models, but the CNN model has too many layers, although there are operations to reduce parameters in each layer structure, the multi-layer data still can lead to the problem of excessive calculation and low efficiency. So in this paper, an efficiency optimization scheme for CNN Chinese acoustic model is proposed.

### 3. CNN acoustic model efficiency optimization scheme

The spatial complexity parameters of convolutional neural networks mainly include feature map size, number of neural network parameters and convolution process. The computational complexity of the convolution process accounts for the major complexity of the entire CNN space. While the CNN acoustic model using traditional convolution process is too cumbersome when dealing with large corpus speech features, which causes low efficiency of the model. In this perspective, this paper proposes a scheme for optimizing the convolution calculation process of CNN acoustic model. Considering that the current computer processor supports SIMD (Single Instruction Multiple Data) stream, this scheme specifies that the processor uses the SIMD based NEON instruction set. Each time the computer reads the data, it usually reads 4 or 8 bits of data. Therefore we generally keep the number of features of the feature map at a multiple of 8 when training the model, which is convenient for later calculation.

#### 3.1 Convolution optimization algorithm

The traditional convolution method is mainly a one-to-one convolution of the feature map and the convolution kernel. It is assumed that there are 16 sets of Feature maps with a size of 3*3, corresponding to 16 sets of kernels with a size of 3*3, each Feature map convolves a kernel. The
feature map and the kernel have the same size, there is no need to expand the edge, so \( p = 0 \), the offset value is not considered here, then there are 16 sets of 1*1 Feature maps as the output. The convolution calculation here is equivalent to multiplying and adding the corresponding positions between the matrices. The specific convolution calculation process is shown in Figure 2.

For this example, we use the optimized convolution method to transform a single 3*3 Feature map into 1*9, and then integrate all Feature maps into one large image. Similarly, all the kernels are also integrated into one big picture. At this time, the feature map becomes 9*16 and kernel becomes 16*9. Considering that the computer can read 8-bit data at a time, we can use the way of Figure 3 to read 8 sets of data into the buffer area each time during the convolution process, and then combine the SIMD instruction parallel operation to improve the efficiency of the model. We use convolution from top to bottom for the input feature map and convolution from left to right for kernel. The algorithm sets the loop until the entire Feature map convolution is completed and outputs 1*16 output feature map.

In practical applications, the size of the Feature map may be different from kernel. At this time, we can set different loop to read Feature map data. For example, 16 sets of 6*6 Feature map convolved 16 sets of 3*3 kernels, the data should be read in a loop as shown in Figure 4, repeating similar steps until the convolution is completed. In practical applications, the specific loop mode should be set according to the feature map and the kernel size. When the convolution cannot be just convolved, the expansion of the edge can be considered to complete the convolution. Generally, the extended edge is supplemented with 0. The specific convolution process is not described here.

### 4. Comparative experiments and results analysis

#### 4.1 Unit experiment
4.1.1 Voice database

The purpose of this experiment was to compare the efficiency differences between the CNN acoustic model using traditional convolution methods and optimized convolution methods. The experimental speech database comes from IFLYTEK's Chinese input method voice dictation task data set, it contains about 6 hours of audio data, which is divided into 6000 sentences and 57824 words, and the size is 1.4GB. Table 1 lists the information for the voice database. The training set is used for acoustic model training, the development set is used for model parameter selection, and the test set is used for model efficiency testing. The main characteristic variables in the experiment are the 40-dimensional Mel_filter banks and the experimental CNN acoustic model has 43 layers.

| Voice database   | Number of speakers | Men:women | Number of sentences | Duration(hours) |
|------------------|--------------------|-----------|---------------------|-----------------|
| Training set     | 192                | 103:89    | 3000                | 3.2             |
| Development set  | 95                 | 45:50     | 600                 | 0.6             |
| Test set         | 55                 | 23:22     | 2400                | 2.2             |

Table 1. Statistics of the voice database

4.1.2 Unit experiment results and analysis

The experiment used 8bit fixed-point data and SIMD instruction set to compare the response time difference by changing the number and size of the feature map. The experiment was completed on a Windows server and the car machine provided by Speech recognition group of IFLYTEK Company is used as the main experimental equipment. The device configuration information is shown in Table 2. The specific experimental data is shown in Table 3.

| Software          | System name and core numbers | android32, 4core | CPU frequency | 1Ghz |
|-------------------|------------------------------|------------------|---------------|------|
| Hardware          | i.MX6Q Board                 |                  |               |      |

Table 2. Information table of experimental equipment

| Number of feature maps | Feature map width : height | Traditional model response time (ms) | Optimized model response time | Optimized model response time / Traditional model response time(%) |
|------------------------|---------------------------|-------------------------------------|------------------------------|-------------------------------------------------------------------|
| 32                     | 122:5                     | 6036                                | 1030                         | 17.06                                                             |
|                        | 372:8                     | 74522                               | 12483                        | 16.75                                                             |
|                        | 472:7                     | 39394                               | 6793                         | 17.24                                                             |
| 64                     | 172:7                     | 28566                               | 4719                         | 16.52                                                             |
|                        | 372:8                     | 74522                               | 12483                        | 16.75                                                             |
|                        | 472:7                     | 78840                               | 13247                        | 16.80                                                             |
| 128                    | 72:6                      | 18861                               | 3024                         | 16.03                                                             |
|                        | 122:5                     | 24215                               | 3882                         | 16.03                                                             |
|                        | 122:5                     | 40413                               | 6485                         | 16.05                                                             |
| average value          | 255.33:6.67               | 42818.78                            | 7127.33                      | 16.58                                                             |

According to the experimental data, it can be clearly concluded that the average response time of the optimized acoustic model is only about 16.58% of the traditional model. Using the optimized convolution scheme can greatly improve the efficiency of the acoustic model.

4.2 System experiment
4.2.1 Language model and speech recognition criteria

In order to verify that the optimized acoustic model has no impact on the speech recognition accuracy, we applied the two trained acoustic models to the speech recognition system separately. The experiment used the second and third order n-gram statistical models based on the main vowel modeling. The text corpus of the training language model comes from reading Chinese novels. The environment for reading is the office and the street and the reader is 400 students. The sampling frequency of voice data is 16KHZ, and the sampling bit is 16bit. The parameters of the text corpus are shown in Table 4.

| Corpus     | Number of sentences | Number of words | Word prime number | Syllable number |
|------------|---------------------|----------------|-------------------|-----------------|
| Training set | 984054              | 5880031        | 11762019          | 17840831        |

The evaluation standard of speech recognition is as follows: We suppose the number of all the words in the test data is N, the number of inserted error words is I and the number of correctly recognized words is H, then the correct rate of word recognition is equal to \( \frac{H - I}{N} \). The test set for verifying the correct rate of speech recognition comes from the IFLYTEK Chinese input method speech dictation data set, The data size is 500MB and the duration is 2.5 hours. It contains 2000 sentences and 23800 words.

4.2.2 System experiment results and analysis

The experiment was done on a Linux server. The server information is shown in Table 5, and the specific experimental data is shown in Table 6.

| Hardware            | processor       | Intel(R) Xeon(R)CPU E5-2650 v4 @ 2.00GHz*32core |
|---------------------|-----------------|-------------------------------------------------|
| RAM                 | 512G            |

Table 6. Comparison of recognition accuracy

| Acoustic model      | Word correct rate (%) |
|---------------------|-----------------------|
| Traditional acoustic model | 78.54     |
| Optimized acoustic model     | 78.54     |

It can be seen from the experimental results that two acoustic models are deployed on the speech recognition system, the recognition accuracy rate does not fluctuate. Because the optimization scheme only optimizes the convolution method of the convolutional layer, the convolution result doesn’t change. Therefore, the optimization scheme does not adversely affect the correct recognition rate of the entire speech recognition system.

5. Summary

In this paper, to solve the problem of low efficiency of CNN Chinese acoustic model when processing a large number of feature maps, we propose a method of optimizing convolution operation. The new method improves efficiency by centrally processing the input feature map and the convolution kernel, separately expending all the feature maps and convolution kernels into large feature maps and a convolution kernels, and using the SIMD instruction set. Experiments show that the optimization scheme can improve the efficiency of the whole system while ensuring speech recognition accuracy. In addition to the optimization scheme in this article, there are still many excellent optimization schemes, and we will continue to learn and pay attention to CNN acoustic model related technologies\(^3\) in the future work.

References

[1] Abdel-Hamid O, Mohamed A, Jiang H, et al. (2014) Convolutional neural networks for speech
recognition. J. IEEE Transaction on Audio, Speech, and Language Processing, 22 (10): 1533-1545.

[2] Krizhevsky.A, Sutskever.I, Hinton.G.E.(2012) ImageNet Classification With Deep Convolution Neural Network. In: Neural Information Processing Systems Conference. Lake Tahoe, Nevada, USA. pp. 1106-1114.

[3] Hamid.O.A, Deng.L. Yu.D. (2013) Exploring Convolutional Neural Network Structure and Optimization Techniques for Speech Recognition. In: INTERSPEECH. Lyon, France. pp. 3366-3370.

[4] Mohamed.A, Dahl.G, Hinton.G. (2012) Acoustic Modeling Using Deep Belief Networks. J.IEEE Transaction on Audio, Speech, and Language Processing, 1:14-22.

[5] Hinton.C, Deng.L, Dong.Y, et al. (2012) Deep neural network for acoustic modeling in speech recognition: The shared views of four research group. J. Signal Processing Magazine.29(6). pp. 82-97.

[6] Sainath.T.N, Mohand.A, Kingshury.B, Ramabhadran.B.(2013) Deep Convolutional Neural Networks for LVCSR. In: ICASSP. Vancouver, Canada. pp. 8614-8618.