EAI Endorsed Transactions
on e-Learning

Research Article EAI.EU

Six-layer Optimized Convolutional Neural Network for Lip Language Identification

Yifei Qiao[1], Hongli Chen[1], Xi Huang[1], Juan Lei[1], Xiangyu Cheng[1], Huibao Huang[1], Jinghan Wu[1], Xianwei Jiang[1,*]

[1]School of Mathematics and Information Science, Nanjing Normal University of Special Education, Nanjing 210038, China

Abstract

INTRODUCTION: Lip language is one of the most important communication methods in social life for people with hearing impairment and impaired expression ability. This communication method relies on visual recognition to understand the meaning expressed in communication.

OBJECTIVES: In order to improve the accuracy of this natural language recognition, we propose six-layer optimized convolutional neural network for lip recognition.

METHODS: The calculation method of the convolutional layer in the CNN model is used, and two pooling methods are compared: the maximum pooling operation and the average pooling operation to analyse the most important feature data in the picture. In order to reduce the simulation in the model training process, the closing rate has been optimized by introducing Dropout technology.

RESULTS: It shows that the recognition accuracy rate based on the six-layer convolutional neural network can reach 85.74% on average. This method can effectively recognize lip language.

CONCLUSION: We propose a six-layer optimized convolutional neural network method for lip language recognition, and the identification of lip language features of this method is better than 3D+ DenseNet +1 × 1 Conv +resBi-LSTM, 3D+CNN, ConvNet+2 -256-LSTM+VGG-16 three advanced methods.

Keywords: lip language identification, convolutional neural network, Batch Normalization, dropout

Received on 20 June 2021, accepted on 11 August 2021, published on 20 August 2021

doi: 10.4108/eai.20-8-2021.170751

*Corresponding author. Email: jxw@njts.edu.cn

1. Introduction

Human perception of speech is a very complex multi-modal process, in addition to acoustic features, it also includes the comprehensive application of features, grammar, semantics, and contextual knowledge. A large number of studies have shown that both acoustic language and visual language play a very important role when people understand the content of speech. In terms of the production mechanism of speech, both acoustic speech and visual speech are produced by the vocal organs such as vocal cords, soft palate, tongue, teeth, lips, jaw and nasal cavity [1]. Some of the articulators can be seen, so there is a certain internal connection between acoustic speech and visual speech. But not all internal organs are visible, and there is not a simple one-to-one relationship between visual speech and acoustic speech. In daily life, people's understanding of speech depends not only on hearing but also on vision. For some people with hearing impairment, spoken language recognition is an important way for them to obtain external voice information. Among the speech organs, only the lips and the lower jaw are completely visible, and the tongue and teeth are partially visible. These organs carry the visual characteristics of the speech. Lip reading mainly judges the speaker's content through the change of mouth shape, and the research of lip reading belongs to the category of human-computer interaction. At this stage, the research on lip reading is mainly based on the recognition of speaker content, and effective pre-processing (including video cutting, image enhancement, lip edge positioning) of the collected speaker lip motion image sequence, and how to select appropriate features after pre-processing is To solve the key problem of recognition accuracy, in the study of lip reading recognition, the shape feature and image feature based on the sequence of mouth shape changes are used for experiments [2].

Since deep learning is essentially a data-driven algorithm, more and more successful deep learning examples show that the quality of the data set determines the quality of the deep learning algorithm model training results, and the recognition of lip language is no exception. The current mainstream method is to perform feature segmentation on the lip shape after in-depth analysis. After multiple training and selecting appropriate thresholds based on experience, a better lip segmentation binary image is obtained; after obtaining the lip shape binary image, construct The template smooths the image and extracts the edges, and selects an appropriate number of edge feature points; finally, the neural network is applied to repeatedly train the edge feature points to obtain a smooth edge fitting curve. The most advanced lip language recognition technology so far is an automatic lip language tagging system based on the Pyramid LK (Lucas-Kanade) optical flow method. The system first uses voice processing technology and facial lip region positioning technology to pre-process the video, and then uses optical flow method to calculate the movement information of the lips between adjacent frames to accurately mark the time corresponding to the lips change. Mark the task. Compared with the method of labelling by speech recognition alone, the lip samples labelled by this system are more accurate and the quality of the data set is higher. In order to realize the recognition of Chinese lip language, a Chinese Phrase Lip Data Set (CPLDS) was established using this system. In the construction of the deep learning model for lip recognition, starting from the characteristics of lip movement, because lip recognition not only needs to identify the information of the lip region picture space, but also need to pay attention to the relationship of the picture sequence over time. Use the improved VGG (Visual Geometry Group) convolutional neural network to extract the spatial features of the lip pictures, and then use the GRU (Gated Recurrent Unit) recurrent neural network to extract the temporal features of the lip movement, and finally combine the two to construct the lip language. Identify deep learning models. In the design of the loss function, CTC (Connectionist Temporal Classification) is used as the timing output loss. In the training process of the deep neural network model, transfer learning is used to improve the generalization ability of the model. At the same time, the batch normalization BN (Batch Normalization) and the dropout method (Dropout) are used to prevent the model from overfitting[3].

In this method of lip language recognition, the network model can be decomposed into two parts, an image model and a language model, both of which require a large amount of sample information. Compared with other languages that

only consist of letters, Chinese characters are more complicated. There are more than 1,000 pronunciations of Hanyu Pinyin, and the number of Chinese characters exceeds more than 9,000. These reasons make the recognition of Chinese lip language even more effective. difficult. Therefore, in general lip language recognition, the correct rate will be higher and higher as the number of samples increases, but there is still a long way to go to achieve a high correct rate like speech recognition[4].

Convolutional neural network is a kind of feedforward neural network developed in recent years, that is, the simplest neural network, and a college recognition method that can attract wide attention. Convolutional neural network is a feedforward multi-layer network. Information flows in only one direction, that is, from input to output. Each layer uses a set of convolution kernels to perform multiple transformations. The CNN model mainly includes a convolutional layer, a pooling layer, and a fully connected layer[5]. His neurons are arranged hierarchically, and each neuron is only connected to the neuron of the previous layer, receives the output of the previous layer, and outputs it to the next layer. His artificial nerve can respond to the surrounding units partially covering a certain range, and has great advantages in processing large images. The network avoids the pre-processing of complex images and can directly input the original image. Based on the CNN model, combining multi-layer convolution and multi-layer pooling to generate a new network model can improve the accuracy of the network structure[5].Based on the results of previous researchers by large-scale computing cluster, dedicated hardware and vast amounts of data, convolution neural network in image classification and object recognition has been widely and useful applications. Although they do not have the creativity of human mind, strong ability of object recognition is worth our using for reference.

2. Dataset

The pronunciation of Chinese characters is made up of pinyin, and pinyin is made up of syllables and tones. Since 1955, Chinese pinyin has been used as a tool to assist the pronunciation of Chinese characters. It is similar to the phonetic alphabet of English, but it is very different[6]. Research on Chinese shows that the pronunciation of Chinese characters can be represented by more than 1,300 syllables. A syllable is composed of initials and finals. The initials are the beginning of the entire syllable, and the rest are finals. There are 23 initials, which can be divided into bilabial sounds, Labiodental, alveolar, gingival palatal, tongue curl and velar[7].The pronunciation classification is shown in Table 1. There are 32 Chinese phonemes in total, as shown in Table 2.

Table 1 Pronunciation classification

Bilabial	Labiodental	Alveolar	Gingival palatal	Retroflex	Velar
b	f	d	j	zh	g
p		t	q	ch	k
m		n	x	sh	h
		l		r	
		z			
		c			
		s			

Table 2 32-Chinese phonemes

Consonant	Gingival palatal
b, p, m, f,d, t, n, l,g,k,h,j, q, x,	a, o, e,i, u, ü, ê,
zh, ch, sh, r,z, c,s,ng	-i[before], -i[after],er

Different pronunciation parts and pronunciation methods determine the differences in sounds, but when there is no sound and the pronunciation is judged only by vision, it is difficult to distinguish certain phonemes. Unlike most image recognition tasks, the pinyin sequence recognition requires the network to be able to capture the most subtle features of the image, especially the movement changes between the lips.

On this basis, we collected experimental models of lip language (see Figure 1). The lip shapes of some letters are basically similar. For example, the visual effects of c, e, p, q, s, t, x, z are similar, and the corners of the mouth are flat and the teeth are pronounced; b, d, o, and u are round and toothless mouth shapes; a, f, h, i, l, m, n, r, w, x read history when several separate lips flat upper and lower teeth; J lip bite sound bite the lower lip of the teeth, v is from two lips closed slit tonguing.

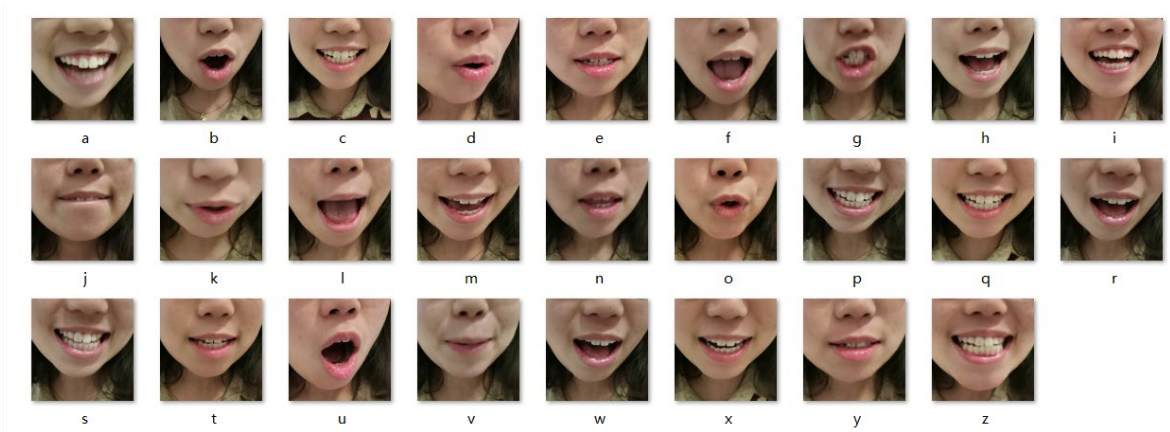


Figure 1. Collected lip language patterns

3. Methodology

3.1 Convolutional Neural Network

Generally speaking CNN convolution neural network recognition images, typically take steps for, first by the local characteristics of convolution layer is responsible for extracting image; Second pooling layer used to greatly reduce parameter magnitude (d); Finally the connection layer similar part of the traditional neural network, is used to output the desired results based on the CNN model, multiple network models have been proposed and applied to speech recognition and image recognition, such as convolutional neural networks[8]. Image recognition is not an easy task, a good approach is to apply metadata to unstructured data, one way to solve this problem is the use of convolution neural network.

3.2 Convolutional Layer

Convolutional layer in convolutional neural network, each convolution layer is composed of several convolution units, and the parameters of each convolution unit are optimized

by back propagation algorithm. The purpose of convolution operation is to extract different features of input. The first convolution layer can only extract some low-level features, such as edge, line and angle. More layers of network can extract more complex features iteratively from low-level features.

Convolution layer is the most computationally intensive part of convolution neural network. In this layer, the output characteristic graph of the upper layer is convoluted with a trainable convolution kernel to obtain the corresponding results[9].

$$W = [w_1, w_2, \dots, w_n]^T \quad (1)$$

$$X = [x_1, x_2, \dots, x_n]^T \quad (2)$$

$$Y = W^T + b \quad (3)$$

W is the weight vector; X is the input eigengraph vector; B is bias; Y is the output characteristic graph.

In convolutional neural network, the computation of convolution layer takes up most of the computation of the whole network, and the required parameters are relatively small, which is especially suitable for parallel expansion computation. The convolution layer is usually followed by a pooling layer, so this paper implements a pipeline structure design for the convolution layer and pooling layer . The size of the convolution kernel is selected to be 3×3. At the same time, in order to balance the amount of calculation, the size of the input feature map is selected to be 224×224. The pooling layer selects one of every four adjacent pixels as the output feature map pixel. The whole design outputs 1 pixel per clock cycle, and the corresponding pooling layer needs 4 pixels per clock cycle in parallel. Figure 2 shows the hole block diagram[10].

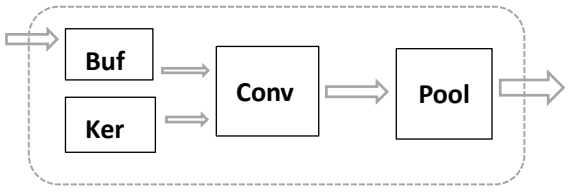


Figure 2. The whole block diagram

Calculation method of convolution layer:

$$F_1 = W_{11} \times R + W_{12} \times G + W_{13} \times B + b_1 F_2 = W_{21} \times R + W_{22} \times G + W_{23} \times B + b_2 F_3 = W_{31} \times R + W_{32} \times G + W_{33} \times B + b_3 F_4 = W_{42} \times G + W_{43} \times B + b_4 F_5 \quad (4)$$

Where F represents each feature graph, W represents convolution kernel, and * represents convolution operator.

3.3 Fully Connected Layer

The connected layer is actually a convolutional operation in which the convolution kernel size is the upper level feature size. The result of convolution is a node, which corresponds to a point in the fully connected layer.

Each node of the fully connected layer is connected with all nodes of the upper layer to synthesize the features extracted from the front. Because of its fully connected characteristics, the parameters of the fully connected layer are usually the most. The specific implementation of convolution is shown in Figure 3.The left side is the 5 * 5 input feature map, the grey part is the 3 * 3 convolution core, and the right side is the output feature map. After each calculation, the convolution core moves one step to the right, usually 1. After each line of convolution, the convolution core returns to the starting position and moves one line down until all convolutions are completed.

3.4 Pooling Layer

The pooling layer, also known as the lower sampling layer, is used to carry out the lower sampling of the input feature map and is generally used between continuous convolutional layers to reduce the number of parameters [11].The goal of the pooling layer is to bring a certain degree of invariance to the changes in the position and proportion of the input data, so as to reduce the phenomenon of overfitting to a certain extent, and carry out feature dimensionality reduction on the image to remove redundant information and retain the most important features, so that the features in the image remain unchanged.

The pooling layer compacts the amount of data and the number of parameters of the image output from the convolutional layer [12].The subsampled convolution kernel only takes the maximum or average value of the corresponding position as the eigenvalue of the point. The operations corresponding to these two values are

respectively Max Pooling and Average Pooling, which are the most used commonly, as shown in Figure 4.Two pooling methods are presented respectively.

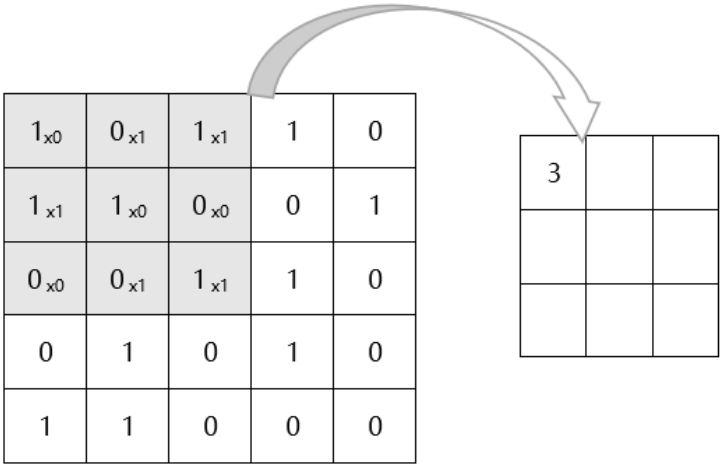


Figure 3. The concrete realization of the convolutional layer

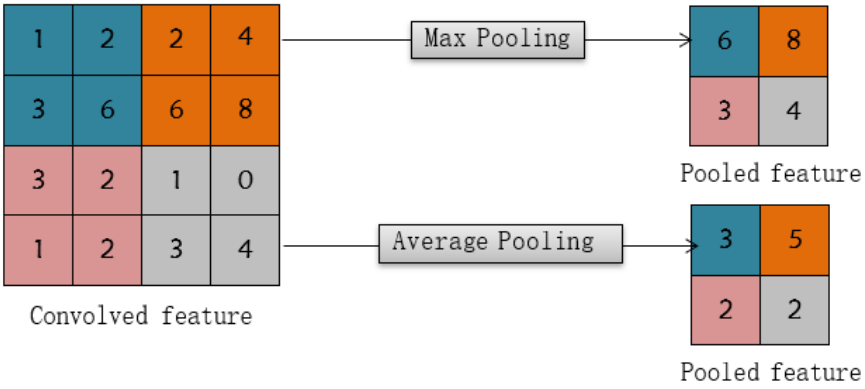


Figure 4. Pooling method

As shown in Figure 4, the convolution kernel size is 2 * 2, under the condition of step length is 2, after the pool size is a quarter before pooling, in pixel level space, an area of the image characteristics in its neighboring areas may also apply to [13], so pooling operation under the condition of without reducing model accuracy, effectively reduce the number of parameters in the network model.

3.5 Dropout

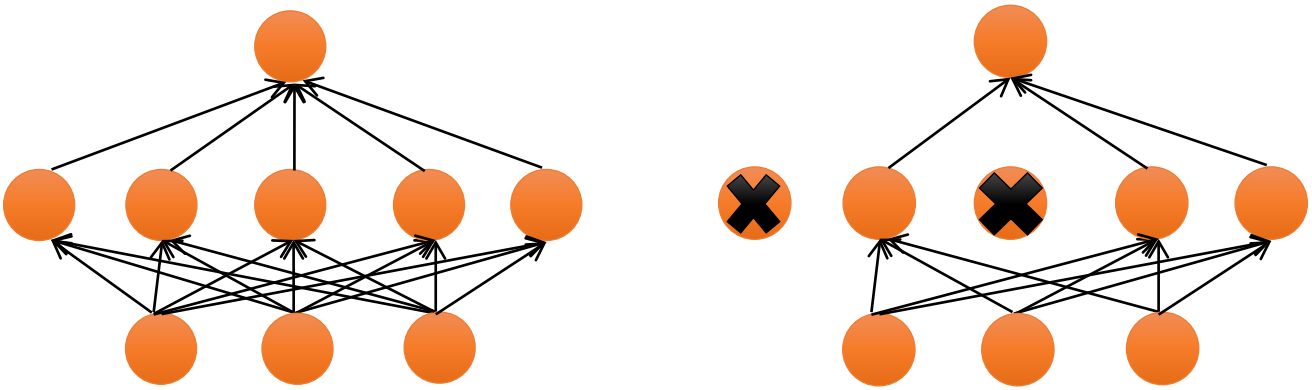
In the training process of deep convolutional neural network, with the increase of network depth, network parameters will become more and more. If the training model has too many parameters and insufficient training samples, the probability of over-fitting phenomenon in the trained model will be very high [13]. It is very small in the training set, but the error is large when the test data is provided to the neural network, which is named as the bad generation of the new data set [14]. To solve this problem,

Dropout was introduced, and Dropout is a method proposed to overcome the overfitting problem. The output of the Dropout layer is.

$$y = k * l(W^T x) \quad (5)$$

Where, $x = [x_1, x_2, \ldots, x_n]^T$ is the input of the full connective layer, $W \in R^{h \times n}$ is a weight matrix, k is a binary matrix of size h and obeys the Bernoulli distribution of parameter p.

Dropout works by "dropping" some neurons each time it passes forward, which means it randomly sets some neurons to zero. Each cell has a fixed probability, p, which is independent of the other cells. Probability p is usually set to 0.5 [15]. In this case, Dropout randomly generates the most network structure, which helps reduce overfitting. Figure 5 is a graphical illustration of Dropout:



Network not using Dropout

Use Dropout's network

Figure 5. Dropout Schematic

4. Experiments Results and Discussions

4.1 Statistical Results

This experiment is based on Window10 system, I5 CPU computer side. This experiment is based on the window10 system, i5CPU computer terminal. The structure of our CNN network is shown in the Figure 6. Block A and Block B are designed in this structure. After data input, Block A is called twice and then Block B is called twice. The features obtained from the previous are continuously integrated through the two layers of Fully Connected Layer, and the Prediction is finally obtained. Block A performs data processing training through convolution, BN (Batch Normalization), and Relu (Rectified Linear Units), and then uses pooling technology to integrate the feature points in the small neighborhood obtained after the convolution layer to obtain new features. The difference between Block B and Block A is that there is one more BN step repetition after Relu.

On this basis, the collected data are tested and analysed by means of the mean iteration algorithm, that is, A gray value T is obtained through iterative calculation, which divides the image into two categories A and B. Satisfies the condition: the mean of class A and class B, and then the mean is exactly equal to T. In these ten experiments, CNN and other advanced algorithms were adopted. After running 10 times, the highest value, lowest value and mean value were obtained. The average accuracy was 85.74%, as shown in Table 3, indicating a relatively high overall accuracy.

Table 3 Statistical result (using dropout)

Serial number	Precision
1	82.81%
2	83.59%
3	84.77%
4	85.55%
5	87.11%
6	83.20%
7	85.16%
8	88.67%
9	88.28%
10	88.28%
Average	85.74%

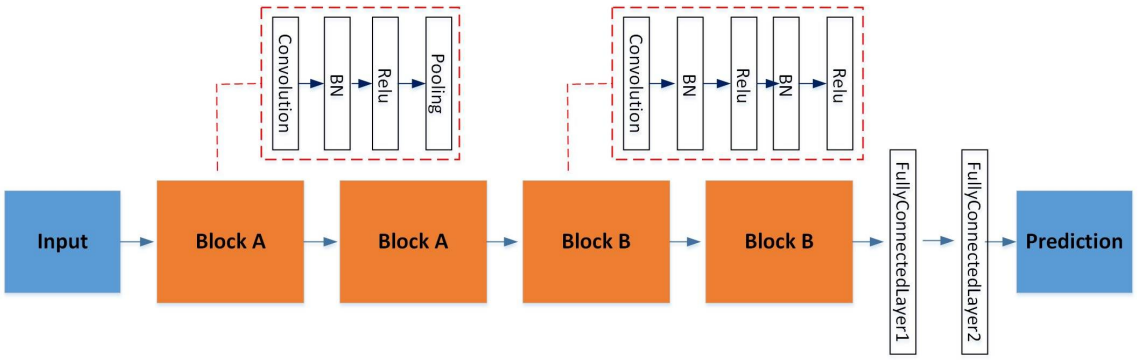


Figure 6. Our CNN network structure

4.2 Optimized algorithm comparison

In the convolutional network, Pooling layers are usually added between the convolutional layers， which is a down-sampling operation. The pooling method imitates the human visual system for dimension reduction (sampling reduction), and represents the image features with a higher level of abstraction. This part of the content is the content on a line, from Hubelle & Wiesel's visual nerve research to Fukushima's proposal, and then to LeCun's Lenet5, which first adopted and used BP for solution. The original impetus is actually biomimetic, building artificial networks that mimic real neural networks[16]. The Pooling can reduce the parameters of the feature map, improve the calculation speed, increase the receptive field, and make the model pay more attention to the global features rather than the location of local features. In this way, it can retain some important feature information, improve the fault tolerance, and prevent over fitting to a certain extent [17].

Deep neural networks have very strong learning abilities and very complex functions that are difficult for humans to understand[14]. In machine learning some of the models, if too many model parameters, and the training sample is too little, it is easy to produce phenomenon of excessive fitting, fitting is a lot of deep learning and the common fault of the machine learning algorithm, embodied in higher prediction accuracy in the training set, and accuracy on the test set have fallen sharply, in order to prevent this kind of phenomenon, namely USES the dropout layer. Dropout is a

common way to curb overadaptation in deep learning. This method was proposed by Hinton et al. (2012), which randomly deletes some neurons in the process of neural network learning. During the training, some neurons were randomly selected and their output was set to 0. These neurons do not transmit signals to the outside[18], which can effectively alleviate the occurrence of overfitting phenomenon and play a certain regularization effect. Dropout can alleviate the fitting problem during the experiment. The following table is to analyse the data without using dropout. Over-fitting will occur when the training data is small. However, the accuracy of Precision is improved by using dropout.

Table 4 Comparison of Precision (no dropout) and Precision (with dropout)

Precision(No dropout)	Precision
77.73%	82.81%
80.86%	83.59%
80.47%	84.77%
76.56%	85.55%
83.59%	87.11%
80.86%	83.20%
82.03%	85.16%
82.81%	88.67%
83.98%	88.28%
82.42%	88.28%
Average：	
81.13%	85.74%

4.3 Comparison to State-of-the-art Approaches

Compared with the most advanced methods in the field of lip recognition, the six-layer optimized convolutional neural network makes the training set and the test set more accurate by comparing the optimization algorithms, and reduces the lack of training samples and possible overfitting. The probability of the phenomenon [13]. At the same time, the six-layer optimized convolutional neural network not only has the good adaptability and autonomous learning ability of traditional neural networks, but also can automatically extract a variety of image features and share weights[19]. We also compared and referred to experiments of the same type. The first experiment designed a Pinyin sequence recognition system that combines three-dimensional convolution, DenseNet and resBi-LSTM. Experiments were performed on the data set NSTDB, and the same in the data processing process Only the part of the lip contour is intercepted to minimize the part of the face. The experiment shows that the pinyin error rate is 50.44%[20].The second type of lip recognition using three-dimensional convolutional neural network is in the test phase. A total of 50 videos are input. There are 50 sentences and 300 words, of which all words in 21 sentences are recognized correctly, that is, the accuracy of the sentence-level network model is 42.0%. A total of 203 words out of 300 words are correctly recognized. The accuracy rate of the word-level network model is 67.7% [21].The third experiment adopts the P2P network model, and its best composition is the ConvNet network and the 2-256-LSTM network. Using different learning rates for training, the highest recognition rate of image features is only 41.28% [22], and the extraction rate of lip language features is not very high in several methods. In these three experiments, the optimized convolutional neural network is not used in depth in the recognition of lip language. In the subsequent experiments, various methods need to be used to minimize the error. In the process of collecting lip language pictures, there are errors due to the different picture information obtained due to the size, angle, and light intensity. The six-layer optimized convolutional neural network has the characteristics of high fault tolerance, which will affect subsequent experiments. Provide great help.

Table 5 Comparison of 3D+ DenseNet +1 × 1Conv+resBi-LSTM,3D+CNN,ConvNet+2-256-LSTM+ VGG-16,Six-layer-CNN

Comparative approach	Recognition accuracy
3D+ DenseNet +1 × 1 Conv	50.44%

+resBi-LSTM	
3D+CNN	67.7%
ConvNet+2-256-LSTM+ VGG-16	41.28%
six-layer CNN (ours)	85.74%

6. Conclusions

On the basis of a lot of research on lip language recognition, we propose a six-layer optimized convolutional neural network to recognize lip language. This article first introduces what lip language recognition is, and introduces the most advanced method in the field of lip language recognition, the shortcomings of the new method, and the three parts of CNN that we use. The second part of the article shows the experimental data set, and then the third part introduces the convolutional neural network model in detail. In the fourth part, the data parameters obtained by the experimental analysis of the model are statistically analysed, optimized algorithm comparison and comparison with the most advanced methods. The average value of the experimental results of this method is 85.74%. The experimental result shows that this method can effectively improve the accuracy of lip recognition and achieve better recognition results. This method can be used for more exploration and application in the future.

Acknowledgements.

This work was supported by National Philosophy and Social Sciences Foundation (20BTQ065), Natural Science Foundation of Jiangsu Higher Education Institutions of China (19KJA310002), The Philosophy and Social Science Research Foundation Project of Universities of Jiangsu Province (2017SJB0668).

References

[1] R. Chuanzhen, Master, Shandong University, 2010, PP.62.

[2] X. Minghui and Y. Hongxun, "is based on sentence-level lip recognition technology Lip-reading Recognition on Sentence," (in chi), *Computer Engineering and Application,* vol. 41, no. 8, PP.86-88, 2005.

[3] W. Youda, "Research on Lip Language Recognition Technology Based on Deep Learning," Master, University of Chinese Academy of Sciences (Xi'an Institute of Optics and Fine Mechanics, Chinese Academy of Sciences), 2019, PP.79.

[4] W. Dan, "Research and Design of Lip Language Recognition Based on 3D Convolution," Master, University of Electronic Science and Technology of China, 2019, PP.70.

[5] G. Rongli, C. Jianrong, W. Shiyu, C. Yan, and C. Na, *small microcomputer system,* PP. 1-6.

[6] Z. Xiaobing, G. Haigang, Y. Fan, and D. Xili, "Research on Chinese Lip Language Recognition Based on End-to-End Sentence Level," *Software Journal,* vol. 31, no. 06, PP. 1747-1760, 2020.

[7] H. Shan, Y. Jiabin, and L. Yaoyao, "Research on Lip Language Recognition Method Based on Visual Features of Chinese Pronunciation," *Computer Engineering and Application,* PP. 1-7.

[8] L. Yanxiao, W. Ping, and S. Qindong, "Image secret sharing scheme based on regional convolutional neural network," *Computer Research and Development,* vol. 58, no. 05, PP. 1065-1074, 2021.

[9] C. Huang, Z. Yongxin, f. plow, W. Hui, and Fengsonglin, "FPGA-based convolutional neural network convolutional layer parallel acceleration structure design," *Microelectronics and Computer,* vol. 35, no. 10, PP. 85-88, 2018.

[10] Z. Yulin, W. Donghui, and W. Leiou, *Network New Media Technology,* vol. 10, no. 01, PP. 47-50, 2021.

[11] L. Shuying, "Compressed sensing reconstruction algorithm based on convolutional neural network," Master, Lanzhou Jiaotong University, 2020,PP.60.

[12] Z. Cheng, C. Jie, and D. Chunyang, "Deep neural network combined with graph convolution for text classification," *Computer Engineering and Application,* PP. 1-12.

[13] J. Tengfei, "Research on Remote Sensing Image Scene Classification Based on Convolutional Neural Network," Master, Henan University, 2019,PP.69.

[14] S. H. Wang *et al.*, "Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling," *Frontiers in Neuroscience,* vol. 12, 2018,PP.11.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems,* vol. 25, no. 2, 2012,PP.9.

[16] A. M. yewuya, "Pooling of deep learning network layer," 2017.

[17] Z. jiaxiaohuo, "Basic composition of neural network-pooling layer, dropout layer, BN layer, fully connected layer 13," 2020.

[18] H. Mengjiao, "Research and implementation of dropout method based on selective area drop," Master, University of Electronic Science and Technology of China, 2020,PP.76.

[19] L. Yuanyuan, "Convolutional Neural Network Optimization and Its Application in Image Recognition," Master, Shenyang University of Technology, 2016,PP.66.

[20] C. Xuejuan, "Chinese lip recognition and keyword detection based on deep learning," Master, Huaqiao University, PP.69, 2020.

[21] J. Haiyang, "Research on Lip Language Recognition Technology Based on Convolutional Neural Network," Master, North China University of Technology, 2020,PP.60.

[22] Y. Fan, "Research and implementation of lip recognition application based on deep learning," Master, University of Electronic Science and Technology of China, 2018,PP.80.