Research Article

A Chinese Lip-Reading System Based on Convolutional Block Attention Module

Yuanyao Lu ¹, Qi Xiao,¹ and Haiyang Jiang²

¹School of Information Science and Technology, North China University of Technology, Beijing, China

²School of Electrical and Control Engineering, North China University of Technology, Beijing, China

Correspondence should be addressed to Yuanyao Lu; luyy@ncut.edu.cn

Received 15 October 2021; Accepted 30 November 2021; Published 23 December 2021

Academic Editor: Jude Hemanth

Copyright © 2021 Yuanyao Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, deep learning has already been applied to English lip-reading. However, Chinese lip-reading starts late and lacks relevant dataset, and the recognition accuracy is not ideal. Therefore, this paper proposes a new hybrid neural network model to establish a Chinese lip-reading system. In this paper, we integrate the attention mechanism into both CNN and RNN. Specifically, we add the convolutional block attention module (CBAM) to the ResNet50 neural network, which enhances its ability to capture the small differences among the mouth patterns of similarly pronounced words in Chinese, improving the performance of feature extraction in the convolution process. We also add the time attention mechanism to the GRU neural network, which helps to extract the features among consecutive lip motion images. Considering the effects of the moments before and after on the current moment in the lip-reading process, we assign more weights to the key frames, which makes the features more representative. We further validate our model through experiments on our self-built dataset. Our experiments show that using convolutional block attention module (CBAM) in the Chinese lip-reading model can accurately recognize Chinese numbers 0–9 and some frequently used Chinese words. Compared with other lip-reading systems, our system has better performance and higher recognition accuracy.

1. Introduction

Lip-reading is a human-computer interaction technology based on AI [1] that has been widely used. In some noisy environments, speech recognition does not work well; lip-reading can use visual signal to enhance the performance of automatic speech recognition. In recent years, lip-reading has been widely used in many fields, such as assisted autonomous driving, assisted hearing for the hearing impaired, and assisted face recognition [2–5].

In traditional lip-reading methods, Sumby and Pollack [6] first took the continuous movement of the lips as visual information to identify verbal information. Then, Petajan [7] firstly proposed the concept of lip-reading system. Based on his theory, Goldeschen [8] combined Petajan's work with the Hidden Markov Model and proposed a method of lip-reading using the dynamic characteristics of lips as input to a Markov chain.

In recent years, artificial neural network (ANN) [9] based on deep learning, which is growing increasingly popular, has gradually been introduced into the field of lip-reading. In 2016, Google cooperated with Oxford University and designed the first lip-reading model based on sentence level, named LipNet [10]. After that, lip-reading based on deep learning has been gradually developing. In 2018, Burton et al. [11] and others used CNN and LSTM as a deep learning network for lip-reading to solve the complex speech recognition problem that the HMM network cannot. In 2019, the attention mechanism was introduced in lip-reading for the first time. Lu et al. [12] and others proposed a model using a composite neural network model of CNN-Bi-GRU-Attention, and the final recognition accuracy reached 86.8%. In 2021, Hussein et al. [13] improved this model and proposed an HLR-Net model mainly composed of the Inception, Bi-GRU, and Attention Mechanism. The author also used the

CTC loss function to match the input and output. The recognition accuracy rate reached 92%.

However, these studies are all based on English words or sentences dataset. Both the models and methods are unsuitable for Chinese characters. This is mainly due to the drastically different pronunciation principle and composition method of Chinese characters. In English, a word is spelled by 26 alphabets. However, the pronunciation of Chinese is mainly composed of vowels and consonants, and it also contains four tones. There are about 1,000 types of pinyin and more than 90,000 Chinese characters. On average, each pinyin corresponds to approximately 90 Chinese characters. Therefore, how to extract features from such an information-rich language is a very big challenge for Chinese lip-reading.

In recent years, Chinese lip-reading has been gradually developing. Chen et al. [14] proposed a Chinese lip-reading model of DenseNet-resBi-LSTM by going from pinyin to Chinese characters, and the accuracy of Chinese sentences recognition reached 50%. Similarly, Zhang et al. [15] proposed a Chinese lip-reading model based on ChLipNet and the recognition accuracy of Chinese sentences reached 59%. However, the accuracy in Chinese lip-reading is not ideal enough.

Because the details of Chinese lip pronunciation are not obvious, it is difficult to extract image features. The traditional CNN networks cannot extract all the information from the lip image. Therefore, in order to solve this problem and improve the accuracy of the Chinese lip-reading, in this work, we propose a Chinese lip-reading system based on the convolutional block attention module. This system consists of three parts: First, the ResNet50 network with a convolutional block attention module [16]. This module (CBAM) consists of two parts, the channel attention and the spatial attention. The channel attention is mainly to compress the feature map in the spatial dimension. In this process, we perform average pooling and maximum pooling to aggregate the spatial information of feature maps. In other words, it is to focus on the important part in this image. Similarly, the spatial attention compresses the channel and performs average pooling and maximum pooling in the channel dimensions. This process is to combine the 1-channel feature maps to a 2-channel feature map. It is to focus on the important channel in this image. We introduce this module into the ResNet50 network to improve the performance of feature extraction and the features are more robust and fault-tolerant. Second one is a GRU network with an attention mechanism. Because lip pronunciation is a continuous process, the visual information is expressed in continuous frames. We use GRU network with attention to learn sequential information among frame features. Third, we use Softmax classifier. It is to classify the results and get the recognition output. In order to train and test our system, we build our own dataset which contains Chinese numbers 0–9 and ten frequently used Chinese words. The experiment shows that our system can completely recognize these words, and the average accuracy reaches 93.2%.

Contribution: we have three main contributions.

- (1) The self-built dataset contains Chinese numbers 0–9 and ten frequently used Chinese words.
- (2) We build a Chinese lip-reading system based on the convolutional block attention module. It can well extract the spatial characteristics of a single lip image, the sequential characteristics of multiple consecutive lip motion image sequences, and finally recognize the results.
- (3) We can come to the conclusion that with our own dataset, the Chinese lip-reading system based on the convolutional block attention module can recognize the Chinese numbers 0–9 and ten frequently used Chinese words, and the accuracy is higher than other Chinese lip-reading models.

2. Lip-Reading Model Based on Convolutional Block Attention Module

With the development of lip-reading, deep learning is gradually introduced into this field. Deep convolutional neural networks (CNNs) such as Vgg16 [17], ResNet [18], Mobilenet [19], Alexnet, and deep recurrent neural networks (RNNs) such as LSTM [20] and GRU [18] are gradually used in lip-reading models. To capture the spatial features of the lip motion image and the sequential features of the continuous lip motion images, this paper proposes a new lip-reading system based on the convolutional block attention module. Its structure is shown in Figure 1.

This model includes five parts:

- (1) Input: video preprocessing: it extracts the key frames of the video and then performs face recognition, locates the lip position in each frame, and obtains ten consecutive lip motion image sequences.
- (2) CNN: uses the ResNet50 network with the convolutional block attention module (CBAM) to learn the spatial features of a single image.
- (3) RNN: uses the GRU network to learn the time features from images sequences.
- (4) Attention: sends the output results from the GRU network in (3) to the time attention and distributes the weight of each output in the GRU network.
- (5) Output: sends the output from (4) into the Softmax classifier and outputs the result.

2.1. ResNet50. With the development of deep learning, the CNN has developed rapidly in recent years. The CNN is a fully connected network with multiple layers that can simulate the structure of human brains and directly perform supervised learning and recognition in images. The CNN mainly consists of unit structures such as convolutional layer, pooling layer, fully connected layer, and Softmax classification layer [21].

With the development of CNN, people find that the more layers of CNN, the better the learning performance of

FIGURE 1: Diagram of lip-reading model based on multiple attention mechanisms.

the model. But with the emergence of deeper CNN, people realize that when the number of layers is larger, the performance of CNN has declined. Because of the gradient disappearance and the gradient explosion, it is difficult to converge the neural network model [22, 23]. Therefore, to improve the performance of CNN and reduce these problems, ResNet50 was proposed. The network structure is shown in Figure 2, where the residual block structure is RB1. We find that the output of the convolutional neural network is $H(x) = F(x) + x$, so the mapping of the residual neural network is $F(x) = H(x) - x$; the fitting performance of $F(x)$ is better. The residual structure of

identity mapping can effectively avoid the problems of gradient disappearance and performance degradation during training.

2.2. Convolutional Block Attention Module (CBAM). We aim to improve the feature extraction performance of convolutional neural networks. In 2018, Woo et al. [16] proposed a convolutional block Attention module (CBAM) for convolutional neural networks. This module will distribute attention weights along two independent dimensions of channel and space. The structure is shown in Figure 3.

FIGURE 2: Structure of Resnet50.

FIGURE 3: Structure of convolutional attention mechanism.

Convolutional block attention module (CBAM) mainly contains the channel attention mechanism and the spatial attention mechanism. The channel attention mechanism is shown in Figure 4.

The channel attention mechanism can be seen as the input feature making a maximum pooling and an average pooling in width and length, sharing the weight, and then combining the output features. Finally, the result will be obtained after Sigmoid function. The channel attention mechanism mainly focuses on determining which content

of each input image is more important. We use average pooling to process every pixel on the image and get the feedback. As for the maximum pooling, only the place with the largest response in the image will get the feedback. Therefore, the channel attention mechanism can be expressed as follows:

$$\begin{aligned}
 M_s(F) &= \sigma(\text{MLP}(\text{AvgPool}(F))) + \text{MLP}(\text{MaxPool}(F)) \\
 &= \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{max}^c))),
 \end{aligned}
 \tag{1}$$

FIGURE 4: Channel attention mechanism.

where σ donates Sigmoid function, avg is the average pooling, max is the maximum pooling, and MLP is the weight sharing.

Similarly, the process of the spatial attention mechanism can be regarded as the compression of the channel; the principle is shown in Figure 5. The function of maximum pooling is to extract the maximum value on the channel. The number of extractions is the height multiplied by the width. The function of average pooling is to extract the average value on the channel, and the number of extractions is also the height multiplied by the width. Then, we connect the 1-channel features into 2-channel features. Therefore, the spatial attention mechanism can be expressed as follows:

$$M_s(F) = \sigma(f^{7 \times 7}[(\text{Avgpool}(F)); \text{Maxpool}(F)]) \\ = \sigma(f^{7 \times 7}([F_{\text{avg}}^s; F_{\text{max}}^s])), \quad (2)$$

where σ donates Sigmoid function, avg is the average pooling, max is the maximum pooling, and 7×7 represents the size of the convolution kernel.

In this paper, we apply the convolutional block attention module (CBAM) on the convolution outputs in each block. We integrate the CBAM into the ResNet50 network, where the integrated Res-Block is shown in Figure 6.

2.3. Gated Recurrent Unit GRU. The traditional recurrent neural network (RNN) is good at processing the sequential data, but with the extension of the RNN network, it may be unable to connect to all information which may cause key information loss. So, it cannot solve the long-distance dependence problem and the performance may drop significantly. Due to this shortcoming of the traditional RNN network, we select the GRU network in this paper, which is a variant of the LSTM. It has a simpler structure and better performance than the LSTM neural network. The structure of GRU is shown in Figure 7.

In Figure 7, r_t and z_t represent the update gate and reset gate, respectively, in the GRU network. The update gate represents a degree of change; in other words, it shows the proportion of the state information from the previous moment that can enter the current state, while the reset gate determines how much information will be written into the current candidate state. Given the input x_t , the output of the reset gate is as follows:

$$r_t = \sigma(W_r g[H_{t-1}, x_t]), \quad (3)$$

where σ is the Sigmoid function, W_r is the weight on reset gate, and H_{t-1} is the output state of the hidden layer at the previous moment.

The output of the update gate is as follows:

$$z_t = \sigma(W_z g[H_{t-1}, x_t]), \quad (4)$$

where σ is the Sigmoid function, W_z is the weight on update gate, and H_{t-1} is the output state of the hidden layer at the previous moment.

The output result of the candidate state is as follows:

$$\tilde{H}_t = \tanh(W_{\tilde{H}} g[r_t g[H_{t-1}, x_t]), \quad (5)$$

where $\tanh$ is the activation function, $W_{\tilde{H}}$ is the weight on candidate state, and H_{t-1} is the output state of the hidden layer at the previous moment.

The output state of the hidden layer is as follows:

$$H_t = (1 - z_t) \cdot H_{t-1} + z_t \cdot \tilde{H}_t. \quad (6)$$

The result of the final output layer is

$$y_t = \sigma(W_o \cdot H_t). \quad (7)$$

Therefore, the unique structure of the GRU network means that it is good at extracting time sequence features and easier to calculate and train.

2.4. Time Attention. Zhang et al. [24] introduced the attention mechanism into the LSTM, which significantly improved the classification and recognition efficiency and performance of LSTM networks. And, Burton et al. [11] introduced the attention mechanism into the lip language recognition system. He proposed a GRU with the attention mechanism model, effectively improving the accuracy of lip-reading. Therefore, this paper adds the time attention mechanism to lip-reading; its structure is shown in Figure 8.

In Figure 8, $x_1 - x_n$ is the gate loop control unit in the GRU network; $H_1 - H_n$ is the output of each gate loop control unit in the GRU network, which is inputted to the time attention mechanism model, where the output of the hidden layer u_t is as follows:

$$u_t = \tanh(W_w H_t + b), \quad (8)$$

where $\tanh$ is activation function and W_w represents the weights.

The weight of each GRU network output result a_t is as follows:

FIGURE 5: Spatial attention mechanism.

FIGURE 6: Res-Block with CBAM.

FIGURE 7: The structure of GRU network.

FIGURE 8: Time attention mechanism.

$$a_t = \text{soft max}_t^T u, \quad w_i. \quad (9)$$

The final output V_t of the attention mechanism layer is

$$v_t = \sum a_t H_t. \quad (10)$$

This model helps the network to pay more attention to the key information in the lip sequence which can be assigned with greater weights, improving the accuracy and the speed of recognition.

3. Experiment

3.1. Dataset. Most of the public lip-reading dataset is in English, which renders them unsuitable for our experiments. In order to meet our own requirements for Chinese lip-reading, we built a lip-reading dataset by ourselves. Our dataset was recorded by 50 people in three days. Each person reads the Chinese number words 0–9 and Chinese words “Chi-Fan” (Eat), “Dui-Bu-Qi” (Sorry), “Ni-hao” (Hello), “Pao-Bu” (Run), “Shui-Jiao” (Sleep), “Wan-Shua” (Play), “Xue-Xiao” (School), “zai-jian” (Good-Bye), “zhong-guo” (China), and “zou-lu” (Walk) 10 times in each day. Finally, we obtained a total of 10,000 videos.

To obtain a continuous sequence of lip movements, we begin to process these videos. The process is shown in Figure 9. First, we extract 10 frames from each of the videos. Second, a 68-point face positioning method is used to perform face recognition on these frames. Third, we locate the lip area and then cut the lip portion out to acquire 10 lip images.

FIGURE 9: Production process of dataset.

3.2. Implementation Details and Evaluation Criteria. We implement and train the network in TensorFlow and Keras. We use Adam optimizer with an initial learning rate of 0.001 and batch size of 64 to train the model.

We divide the dataset into 8000 groups of training set and 2000 groups of testing set. This allows us to perform cross-validation to evaluate the generalization ability of the model. We used the DE-H neural network training algorithm proposed by Bangyal et al. [25] to train our network model and selected the back propagation algorithm to calculate the gradient. We train our model for 50 epochs and then test the model.

In this paper, we use the recognition accuracy to evaluate the performance of the model:

$$\text{accuracy} = \frac{\text{correct identified samples}}{\text{total samples}} \times 100\%. \quad (11)$$

We intend to use the loss function as a performance indicator to evaluate the model training processed:

$$\text{Loss} = - \sum_{i=1}^N y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)}). \quad (12)$$

3.3. Experimental Results and Discussion

3.3.1. Model Training. In this section, we conduct experiments on our self-built dataset. There are 20 labels in our dataset, each label is a Chinese word, and in each label there are 500 sets of data, a total of 10,000 sets of sample data; each set contains 10 lip images, a total of 100,000 images. We train our model for 50 epochs and record the loss function in each epoch. Figure 10 depicts the loss curve of each epoch during the training process.

3.3.2. Ablation Studies. In order to analyze the performance of our model and other common CNN-RNN hybrid neural network models, we have built ten kinds of hybrid neural network models for comparison. These models are Vgg16-LSTM-Attention, Vgg16-GRU-Attention, InceptionV3-LSTM-Attention, InceptionV3-GRU-Attention, ResNet50-LSTM-Attention, ResNet50-GRU-Attention, ResNet101-LSTM-Attention, ResNet101-LSTM-Attention, ResNet152-LSTM-Attention, and ResNet152-GRU-Attention.

FIGURE 10: Loss curve.

All the experiments are conducted on our self-built Chinese lip-reading dataset, and the parameter settings for the models are the same.

After training the abovementioned models for 50 epochs, we use the test set for evaluating. Table 1 summarizes the results of the recognition accuracy of the model after 50 epochs.

From Table 1, we can find that after 50 epochs, the Vgg16-LSTM-Attention model has the lowest accuracy of 95.2%, which is 4.1% lower than the model we proposed. We can also find that the accuracy of ResNet101-GRU-Attention model is 99.6%, same as our model. The accuracy rate of ResNet152-GRU-Attention is 99.8%, 0.2% higher than the ResNet50-CBAM-GRU-Attention model we proposed, but the above two models have larger model parameters and require more time for training. Except for these two models, the accuracy of other models is lower than the model we use.

Figure 11 depicts the accuracy curve of the models with the number of epochs between LSTM and GRU. We divide all models into LSTM and GRU, to compare the performance of the CNN part. We can clearly see that the convergence speed of the two ResNet50-RNN-Attention networks using the convolutional attention mechanism (CBAM) is relatively slow, which indicates that this model has a lower probability of overfitting during the training process, and the learning performance is better. We can also find that the accuracy of the two models is improved, which means that the CBAM convolutional attention mechanism can effectively improve the sensitivity of the convolution channel and the convolution position in the convolution process and boost the network's ability of image feature extraction. Comparing the two models with CBAM, the performance of the model using the GRU is higher than the LSTM.

3.3.3. Experiment Evaluation and Analysis

(1) Experiment Evaluation. To further evaluate the performance of the model, we test the data for each label in the test set to obtain the accuracy of the different labels in the model.

TABLE 1: The accuracy of each model.

Model	Prms (M)	Top-1 accuracy (%)
Vgg16 + LSTM + Attention	144.4	95.2
Vgg16 + GRU + Attention	142.4	95.3
InceptionV3 + LSTM + Attention	31.9	98.2
InceptionV3 + GRU + Attention	29.9	99.1
ResNet50 + LSTM + Attention	33.6	98.2
ResNet50 + GRU + Attention	31.6	99.3
ResNet101 + LSTM + Attention	52.6	97.3
ResNet101 + GRU + Attention	50.6	99.6
ResNet152 + LSTM + Attention	68.3	98.4
ResNet152 + GRU + Attention	66.3	99.8
ResNet50 + CBAM + LSTM + Attention	36.1	98.7
ResNet50 + CBAM + GRU + Attention	34.1	99.6

FIGURE 11: Accuracy rate curve in our self-built dataset. (a) A curve of the accuracy rate changing with CNN and the same of the LSTM-Attention. (b) A curve of the accuracy rate changing with CNN and the same of the GRU-Attention.

We use the test set to evaluate the models mentioned in Section 3.3.2 and record the average accuracy of each label. Our dataset consists of labels: “one (Yi)” to “nine (Jiu)”, “Eat (Chi-Fan)”, “Sorry (Dui-Bu-Qi)”, “Hello (Ni-Hao)”, “Running (Pao-Bu)”, “Sleep (Shui-Jiao)”, “Play (Wan-Shua)”, “School (Xue-Xiao)”, “Good-bye (Zai-Jian)”, “China (Zhong-Guo)”, and “Walk (Zou-Lu)”. The test results are shown in Tables 2 and 3.

From Tables 2 and 3, we can notice that compared with the abovementioned models, the ResNet50-CBAM-GRU-Attention model has the best fitting result for these 20 words. We can also find that the ResNet50-RNN-Attention network with the convolutional block attention module (CBAM) has higher accuracy in each word than the models without it. The most significantly improved word is “Si (Four).” When the RNN selects LSTM, the accuracy rate increases by 1%. When the RNN selects GRU, the accuracy rate increases by 6%.

If we select the accuracy rate above 70% as the compliance indicator for each word, the compliance rate of each model is shown in Table 4.

Table 4 shows that the ResNet50-CBAM-GRU-Attention model has an ideal recognition accuracy for 20 words, and the compliance rate can reach 100%, which is the highest one than any other model. This means that this model has stable recognition accuracy, the best fitting ability, and the best performance.

(2) *Analysis*: we analyzed the relevant data experimental results with the abovementioned models.

We proposed a ResNet50-CBAM-GRU-Attention model; after 50 epochs of training, the accuracy reached 99.6%, and the compliance rate reached 100%.

For Vgg16-LSTM-Attention, this model compared to the model we proposed, the accuracy is 95.2%, 4.4% lower than our model, and the compliance rate is 75%, 25% lower than our model.

TABLE 2: Testing accuracy of each model.

Model/ character	Vgg16 + LSTM + Attention (%)	Vgg16 + GRU + Attention (%)	InceptionV3 + LSTM + Attention (%)	InceptionV3 + GRU + Attention (%)	ResNet50 + LSTM + Attention (%)	ResNet50 + GRU + Attention (%)
Ling (Zero)	63.00	62.00	99.70	89.00	99.50	97.00
Yi (One)	85.00	57.00	50.00	99.00	53.00	98.00
Er (Two)	99.00	99.00	99.40	99.00	99.00	99.00
San (Three)	67.00	67.00	99.00	95.00	87.50	99.00
Si (Four)	80.00	87.00	99.00	81.00	72.00	64.00
Wu (Five)	99.00	99.00	99.00	99.00	99.00	99.00
Liu (Six)	46.00	98.00	98.00	98.00	99.00	99.00
Qi (Seven)	67.00	79.00	51.00	78.00	97.00	98.00
Ba (Eight)	99.00	42.00	99.00	99.00	99.00	99.00
Jiu (Nine)	85.00	25.00	75.00	81.00	97.00	96.00
Chi-Fan (Eat)	76.00	21.00	16.00	81.00	77.00	99.00
Dui-Bu-Qi (Sorry)	24.00	98.00	99.00	99.00	60.00	99.00
Ni-Hao (Hello)	99.00	99.00	99.00	99.00	99.00	99.00
Pao-Bu (Run)	97.00	99.00	99.00	97.00	93.00	98.00
Shui-Jiao (Sleep)	80.00	88.00	99.00	97.00	31.00	99.00
Wan-Shua (Play)	96.00	99.00	99.00	99.00	99.00	98.00
Xue-Xiao (School)	97.00	96.00	86.00	53.00	97.00	99.00
Zai-Jian (Good-Bye)	99.00	95.00	28.00	99.00	89.00	99.00
Zhong- Guo(China)	97.00	92.00	99.00	98.00	99.00	99.00
Zou-Lu (Walk)	98.00	94.00	99.00	99.00	96.00	99.00

TABLE 3: Testing accuracy of each model.

Model/ character	ResNet101 + LSTM + Attention (%)	ResNet101 + GRU + Attention (%)	ResNet152 + LSTM + Attention (%)	ResNet152 + GRU + Attention (%)	ResNet50 + CBAM + LSTM + Attention (%)	ResNet50 + CBAM + GRU + Attention (%)
Ling (Zero)	98.00	16.00	85.00	61.00	99.00	97.00
Yi (One)	93.00	99.00	72.00	99.00	56.00	98.00
Er (Two)	99.00	99.00	99.00	99.00	99.00	99.00
San (Three)	99.00	96.00	98.00	99.00	86.00	99.00
Si (Four)	70.00	99.00	92.00	35.00	73.00	70.00
Wu	99.00	99.00	98.00	99.00	99.00	97.00
(Five)	99.00	99.00	99.00	99.00	99.00	98.00
Liu (Six)	95.00	70.00	85.00	76.00	97.00	98.00
Qi (Seven)	98.00	99.00	98.00	99.00	99.00	99.00
Ba (Eight)	63.00	74.00	16.00	99.00	98.00	97.00
Jiu (Nine)	97.00	99.00	60.00	99.00	78.00	99.00
Chi-Fan (Eat)	95.00	99.00	99.00	99.00	63.00	99.00
Dui-Bu- (Sorry)	99.00	99.00	99.00	99.00	99.00	99.00
Ni-Hao (Hello)	99.00	99.00	99.00	99.00	92.00	96.00
Pao-Bu (Run)	99.00	99.00	73.00	56.00	41.00	99.00
Shui-Jiao (Sleep)	99.00	99.00	99.00	99.00	99.00	99.00
Wan- Shua (Play)	99.00	99.00	99.00	98.00	93.00	99.00
Xue-Xiao (School)	45.00	33.00	92.00	99.00	86.00	99.00
Zai-Jian (Good- Bye)	99.00	99.00	99.00	99.00	98.00	99.00
Zhong- Guo (China)	93.00	99.00	88.00	97.00	97.00	99.00
Zou-Lu (Walk)						

TABLE 4: Compliance rate of each model (the accuracy rate is higher than 70%) .

Model	Qualified number	Compliance rate (%)
Vgg16 + LSTM + Attention	15	75
Vgg16 + GRU + Attention	14	70
InceptionV3 + LSTM + Attention	16	80
InceptionV3 + GRU + Attention	19	95
ResNet50 + LSTM + Attention	17	85
ResNet50 + GRU + Attention	19	95
ResNet101 + LSTM + Attention	18	90
ResNet101 + GRU + Attention	18	90
ResNet152 + LSTM + Attention	19	95
ResNet152 + GRU + Attention	17	85
ResNet50 + CBAM + LSTM + Attention	17	85
ResNet50 + CBAM + GRU + Attention	20	100

For Vgg16-GRU-Attention, the accuracy is 95.3%, 4.3% lower than our model, and the compliance rate is 70%, 25% lower than our model.

For InceptionV3-LSTM-Attention, the accuracy is 98.2%, 1.4% point lower than our model, and the compliance rate is 80%, 20% point lower than our model.

For InceptionV3-GRU-Attention, the accuracy is 99.1%, 0.5% point lower than our model, and the compliance rate is 95%, 5% point lower than our model.

For ResNet50-LSTM-Attention, the accuracy is 98.6%, 1% point lower than our model, and the compliance rate is 85%, 15% point lower than our model.

For ResNet50-GRU-Attention, the accuracy is 99.3%, 0.3% point lower than our model, and the compliance rate is 95%, 5% point lower than our model.

For ResNet101-LSTM-Attention, the accuracy is 97.3%, 2.3% point lower than our model, and the compliance rate is 90%, 10% point lower than our model.

For ResNet101-GRU-Attention, the accuracy is 99.6%, the same as our model, and the compliance rate is 90%, 10% point lower than our model.

For ResNet152-LSTM-Attention, the accuracy is 98.4%, 1.2% point lower than our model, and the compliance rate is 95%, 5% point lower than our model.

For ResNet50-GRU-Attention, the accuracy is 99.8%, 0.2% point higher than our model, and the compliance rate is 85%, 15% point lower than our model.

For ResNet50-CBAM-LSTM-Attention, the accuracy is 98.7%, 0.9% point higher than our model, and the compliance rate is 85%, 15% point lower than our model.

According to the experiment results, we can make a conclusion that the model ResNet50-CBAM-GRU-Attention we proposed can fully recognize 20 Chinese words. It has higher accuracy, fewer model parameters, and the accuracy of each word can reach above 70%. Therefore, our model has great potential for practical application. Its stable recognition ability for Chinese words can ensure the performance of the lip-reading system and reduce errors.

4. Conclusion

In this paper, we propose a Chinese lip-reading model based on the convolutional block attention module. This system is

composed of ResNet50, convolutional block attention module (CBAM) [22], GRU, and attention mechanism. We conduct experiments on the self-built Chinese lip-reading dataset. The results show that the Chinese lip-reading system using the ResNet50-CBAM-GRU-Attention model has a higher accuracy of 99.6% and the accuracy of each word in the dataset has reached above 70%. But for long text sample, the recognition accuracy is not satisfactory.

Our future work plan mainly consists of two parts. The first is to expand the dataset. Since the current dataset mainly contains Chinese characters and words, the next step is to increase the number of long texts and long sentences in the dataset. The second part is to improve the performance of the system. We are considering utilizing the CTC loss function as a solution to solve the long-distance dependency problem of long text or sentence-level lip-reading. Therefore, we plan to design a Chinese sentence-level lip-reading system that can solve long-distance dependence.

Data Availability

All data and program included in this study are available upon request to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (61971007 and 61571013).

References

- [1] S. Pervaiz, Z. Ul-Qayyum, W. H. Bangyal, L. Gao, and J. Ahmad, "A Systematic Literature Review on Particle Swarm Optimization Techniques for Medical Diseases Detection," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 5990999, 10 pages, 2021.
- [2] Z. Chengwei and Z. Haotian, "Lip reading using local-adjacent feature extractor and multi-level feature fusion," *Journal of Physics: Conference Series*, vol. 1883, no. 1, 2021.
- [3] D. Kastaniotis, D. Tsourounis, A. Koureleas, and B. Peev, "Lip Reading in Greek Words at Unconstrained Driving scenario,"

- in *Proceedings of the 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, Patras, Greece, July, 2019.
- [4] M. A. Abrar, A. Islam, M. Hassan, and M. Islam, "Deep Lip Reading-A Deep Learning Based Lip-Reading Software for the Hearing Impaired," in *Proceedings of the 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)*, November, 2019.
 - [5] S. Newman and A. Kirby Laura, "Read my lips! Perception of speech in noise by preschool children with autism and the impact of watching the speaker's face," *Journal of Neurodevelopmental Disorders*, vol. 13, no. 1, 2021.
 - [6] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.
 - [7] E. D. Petajan, "Automatic Lip-reading to Enhance Speech Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 40–47, San Francisco, CA, USA, June 1985.
 - [8] A. J. Goldschen, "Continuous optical automatic speech recognition by lip-reading," *Computational Imaging and Vision*, Springer, New York, NY, USA, 1997.
 - [9] W. H. Bangyal and H. Abdul, A. Wael and A. Hashem, A New Initialization Approach in Particle Swarm Optimization for Global Optimization Problems," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 6628889, 17 pages, 2021.
 - [10] Y. M. Assael, B. Shillingford, S. Whiteson, and N. D. Freitas, "LipNet: End-To-End Sentence-Level Lip-reading," in *Proceedings of the ICLR*, San Juan, Puerto Rico, May 2016.
 - [11] J. Burton, D. Frank, and M. Saleh, "The Speaker-independent Lipreading Play-Off; a survey of lipreading machines," in *Proceedings of the 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, Sophia Antipolis, France, December, 2018.
 - [12] H. Lu, X. Liu, and Y. Yin, "A Patent Text Classification Model Based on Multivariate Neural Network Fusion," in *Proceedings of the 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, November, 2019.
 - [13] D. Hussein, D. M. Ibrahim, A. M. Sarhan, and N. M. Elshennawy, "Hlr-Net: A hybrid lip-reading model based on deep convolutional neural networks," *Cmc-Tech Science Press*, vol. 68, no. 2, pp. 1531–1549, 2021.
 - [14] X. Chen, J. Du, and H. Zhang, "Lipreading with DenseNet and resBi-LSTM," *Signal Image and Video Processing*, vol. 14, no. 6, 2020.
 - [15] X. B. Zhang, H. G. gong, and F. yang, "Chinese sentence-level lip reading based on end-to-end model," *Journal of Software*, vol. 31, pp. 1747–1760, 2020.
 - [16] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)2018*, Springer, Cham, Munich, Germany, October, 2018.
 - [17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, ICLR, San Diego, CA, USA, May, 2015.
 - [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June, 2016.
 - [19] A. G. Howard, M. Zhu, B. Chen, and D. ., W. Wang, MobileNets:efficient Convolutional Neural Networks for mobile Vision Applications," 2017, <https://arxiv.org/abs/1704.04861>.
 - [20] Y. Imrana, Y. Xiang, A. Liaqat, and A. R. Zaharawu, "A bi-directional LSTM deep learning approach for intrusion detection," *Expert Systems with Applications*, vol. 185, Article ID 115524, 2021.
 - [21] A. Panthakkan, S. M. Anzar, S. A. Mansoori, and H. Ahmad, "Accurate Prediction of COVID-19 (+) Using AI Deep VGG16 Model," in *Proceedings of the 2020 3rd International Conference on Signal Processing and Information Security (ICSPIS)*, Dubai, UAE, November, 2020.
 - [22] J. H. Lee, Y. J. Kim, Y. W. Kim, and S. Park, "Spotting Malignancies from Gastric Endoscopic Images Using Deep Learning," *Surgical Endoscopy*, vol. 33, no. 2, 2019.
 - [23] A. Gómez-Ríos, S. Tabik, J. Luengo, A. S. M. Shihavuddin, B. Krawczyk, and F. Herreraa, "Towards Highly Accurate Coral Texture Images Classification Using Deep Convolutional Neural Networks and Data Augmentation," *Expert Systems with Applications*, vol. 118, pp. 315–328, 2018.
 - [24] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions," *Neurocomputing*, vol. 410, 2020.
 - [25] W. H. Bangyal, K. Nisar, Ag. A. B. Ag. Ibrahim et al., "Comparative Analysis of low discrepancy sequence-based initialization approaches," *Using Population-Based Algorithms for Solving the Global Optimization Problems*, vol. 11, no. 16, 2021.