applied sciences

MDPI

Article

Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory

Yuanyao Lu * and Hongbo Li

School of Information Science and Technology, North China University of Technology, Beijing 100144, China; leehongbo@foxmail.com

* Correspondence: luyy@ncut.edu.cn

Received: 27 March 2019; Accepted: 11 April 2019; Published: 17 April 2019

Abstract: With the improvement of computer performance, virtual reality (VR) as a new way of visual operation and interaction method gives the automatic lip-reading technology based on visual features broad development prospects. In an immersive VR environment, the user's state can be successfully captured through lip movements, thereby analyzing the user's real-time thinking. Due to complex image processing, hard-to-train classifiers and long-term recognition processes, the traditional lip-reading recognition system is difficult to meet the requirements of practical applications. In this paper, the convolutional neural network (CNN) used to image feature extraction is combined with a recurrent neural network (RNN) based on attention mechanism for automatic lip-reading recognition. Our proposed method for automatic lip-reading recognition can be divided into three steps. Firstly, we extract keyframes from our own established independent database (English pronunciation of numbers from zero to nine by three males and three females). Then, we use the Visual Geometry Group (VGG) network to extract the lip image features. It is found that the image feature extraction results are fault-tolerant and effective. Finally, we compare two lip-reading models: (1) a fusion model with an attention mechanism and (2) a fusion model of two networks. The results show that the accuracy of the proposed model is 88.2% in the test dataset and 84.9% for the contrastive model. Therefore, our proposed method is superior to the traditional lip-reading recognition methods and the general neural networks.

Keywords: virtual reality (VR); self-attention; automatic lip-reading; sensory input; deep learning

1. Introduction

Machine learning methods have had a great impact on social progress in recent years, which promoted the rapid development of artificial intelligence technology and solved many practical problems [1]. Automatic lip-reading technology is one of the important components of human–computer interaction technology and virtual reality (VR) technology. It plays a vital role in human language communication and visual perception. Especially in noisy environments or VR environments, visual signals can remove redundant information, complement speech information, increase the multi-modal input dimension of immersive interaction, reduce the time and workload of human on learning lip language and lip movement, and improve automatic speech recognition ability. It enhances the real experience of immersive VR. Meanwhile, automatic lip-reading technology can be widely used in the VR system [2], information security [3], speech recognition [4] and assisted driving systems [5]. The research of automatic lip-reading involves many fields, such as pattern recognition, computer vision, natural language comprehension and image processing. The contents of the research involve the latest research progress in these fields. Conversely, the study of lip movement

is also a check and development of these theories. Meanwhile, it will also have a profound impact on content-based image compression technology.

Traditional lip-reading systems usually consist of two stages: feature extraction and classification. For the first stage, most previous feature extraction methods use pixel values extracted from the mouth region of interest (ROI) as visual information. Then, the abstract image features are extracted by discrete cosine transform (DCT) [6,7], discrete wavelet transform (DWT) [7] and principal component analysis (PCA) [7,8]. Therefore, the model-based methods, such as active appearance model (AAM) [9] and active shape model (ASM) [10] form non-rigid models and obtain a set of advanced geometric features which has the characteristics of lower dimensionality and stronger robustness. In the second stage, the extracted features are fed into the classifiers of support vector machine (SVM) [11] and hidden Markov model (HMM) [12].

At present, deep learning has made significant progress in the field of computer vision (image representation, target detection, human behavior recognition and video recognition). Therefore, it is an inevitable trend of scientific research to shift the direction of automatic lip-reading technology from the traditional manual feature extraction classification model to the end-to-end deep learning architecture. In recent years, researchers have introduced attention mechanisms on convolutional neural networks (CNN) to focus on areas of interest, and the classification and target detection of images have also achieved great success. For example, a CNN feature extraction method based on attention mechanism proposed by Vinyals et al. [13]. Furthermore, the mechanism of attention can be successfully applied in recurrent neural network (RNN) to find the relationship between the context. Since the changes between video frames of automatic lip-reading are continuous and happen in time series, the researchers use the long short-term memory (LSTM) network [14], which can find hidden association information in time series data such as video, audio and text. A multi-layered neural network structure of cascaded feed-forward layer and LSTM layer is proposed for word-level classification in speaker-based lip-reading.

Considering that lip motion is a continuous process with time information, visual content can be represented by consecutive frames. Therefore, we proposed a hybrid neural network architecture combining CNN and attention-based LSTM to learn the hidden correlation in spatiotemporal information [15] and used the weights of attention to express the importance of keyframes.

Our proposed method for automatic lip-reading recognition can be divided into four parts: Firstly, we extracted keyframes from a sample video, used the key points of the mouth to locate the mouth area to reduce the complexity of redundant information and computational processing in successive frames. Then, features were extracted from the original mouth image using the VGG19 network [16], which consists of 16 convolution layers and three fully connected layers. Thirdly, we used attention-based LSTM network to learn sequential information and attentional weights among video keyframe features. Finally, the final recognition result was predicted by two fully connected layers and a SoftMax layer. The SoftMax function converts predicted results into probability.

The main advantages of our method: (1) The VGG19 is equipped to overcome the image deformation including translation, rotation and distortion. Therefore, the extracted features have strong robustness and fault tolerance. (2) Attention-based LSTM is good at finding and exploiting long time dependencies from sequential data, and the introduction of attention mechanism makes the network selectively focus on active video information and reduce the interference of invalid information. Therefore, the relationship of the features among frames is connected and strengthened.

The rest of the paper is organized as follows: in Section 2, we introduce the preparation work and the architecture of the lip-reading model. Experimental results and analysis of our proposed method are presented in Section 3. Section 4 offers conclusions and suggestions for future research directions.

2. Proposed Lip-Reading Model

In this section, the proposed framework and main steps are discussed in detail according to the following four parts. Firstly, we need to preprocess the dynamic lip videos, including separating

audio and video signals, extracting keyframes and positioning the mouth. Secondly, features are extracted from the preprocessed image dataset by using CNN. Then, we use LSTM with attention mechanism to learn sequence information and attention weights. Finally, the ten-dimensional features are mapped through two fully connected layers, and the result of automatic lip-reading recognition is predicted by SoftMax layer. SoftMax normalizes the output of the fully connected layers and classifies it according to probability. The sum of probabilities is one. Our proposed CNN with attention-based LSTM architecture is shown in Figure 1.

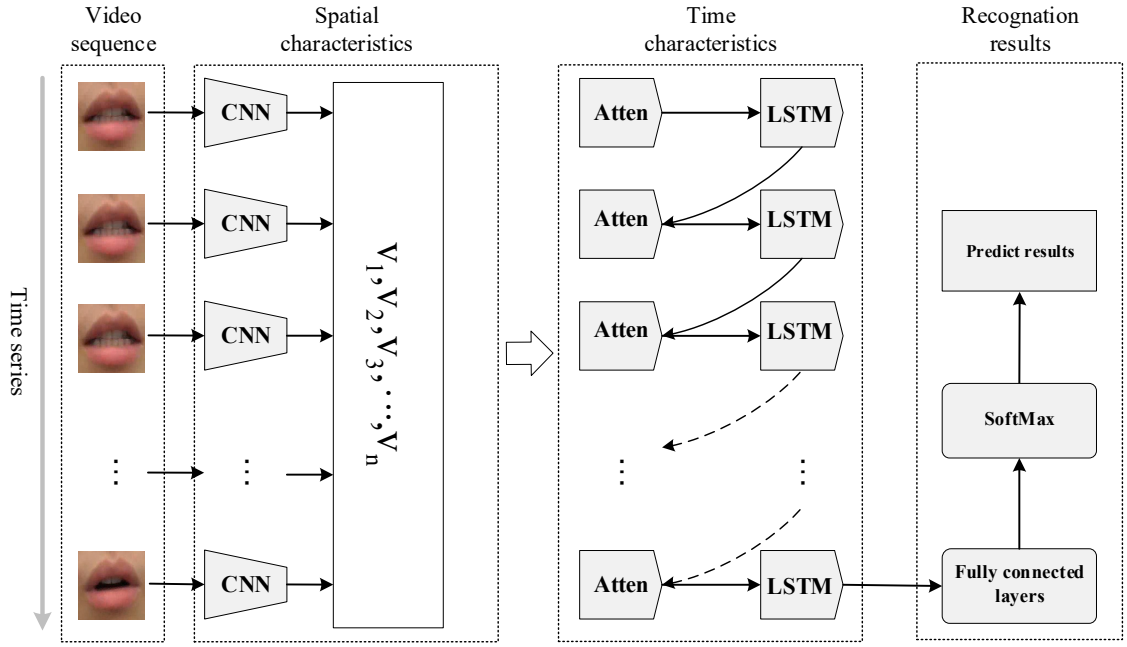


Figure 1. The architecture of our proposed convolutional neural network (CNN) with attention-based long short-term memory (LSTM).

2.1. Video Preprocessing

We preprocess the sequential images of lips in order to balance training speed and recognition results, including keyframes extraction and lip location segmentation [17].

Generally, the data from video capture is about 25 frames per second. Since there is a difference in the length of each utterance and any word actually pronounced has a series of redundant information about the movement of the lips, it is difficult for the model to extract the image features and discover the hidden relationship of the sequences. Therefore, we try to remove redundant information from all the original images, extract keyframes and segment lips in the following four steps as experimental datasets:

- The time of the utterance is divided into 10 equally interval portions, and a random frame of each portion is selected as a keyframe, thus each word obtains a sequence image frame of equal length.
- We use OpenCV library to load images and convert them into a three-dimensional matrix [18]. Then, we use the facial landmark detection of Dlib toolkit [19]. It takes the face images as input, and the returned face structure consists of different landmarks for each specific face attribute. We choose to locate the seven key points of the mouth, labeled as: 49, 51, 53, 55, 57, 58, 59.
- We segment the mouth images and remove the redundant information, then calculate the center position of the mouth based on the coordinate points of the image boundary, denoted as (x_0, y_0). The width and height of the lip image are represented by w and h, respectively, L_1 and L_2 represent

the left and right, upper and lower dividing lines surrounding the mouth, respectively. According to the following formula to calculate the bounding box of the mouth:

$$L_1 = x_0 \pm \frac{w}{2}, \tag{1}$$

$$L_2 = y_0 \pm \frac{h}{2}, \tag{2}$$

- After the mouth segmentation step, the original dataset will be processed into 224 × 224 pixels which take lips as a standard. This method has the characteristics of strong robustness, high computational efficiency and consistency of eigenvectors. The processes of pretreatment are as shown in Figure 2.

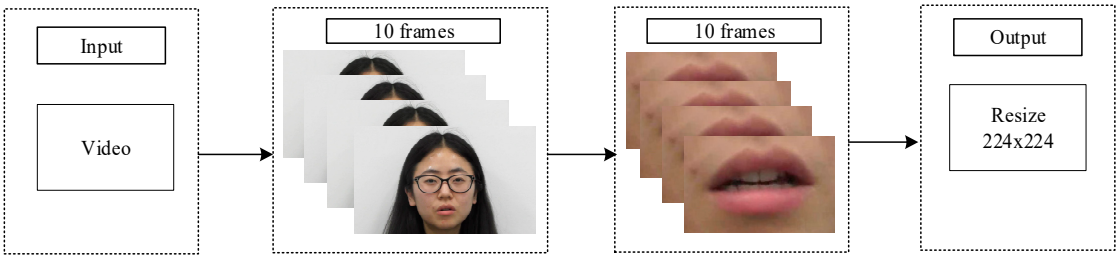


Figure 2. Preprocessing steps with example frames.

2.2. Attention-Based LSTM

In general, RNN can obtain sequence output with time information through sequence input [20]. For example, LSTM network which is a special type of RNN can learn long-term dependency information. LSTM was proposed by Hochreiter and Schmidhuber [21] and it was improved and promoted by Alex Graves in 2012 [22]. In many practical applications, LSTM has achieved considerable success and it has been widely used.

The first step in LSTM is making a decision to discard useless information from the cell state, which is accomplished by a decision called "the forget gate". This gate reads h_{t-1} and x_t, then it outputs a value in the [0,1] interval for each number in cell state C_{t-1}. The calculation process is as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{3}$$

where σ is the hidden activation function, h_{t-1} is the hidden state at time $t-1$, x_t is the input at time t, and b is the bias.

Then, it is determined that new useful information is stored in the cell state. It consists of two parts: First, a sigmoid layer is called the "input gate layer" and it determines which value will be updated. Then a new candidate value vector is created by tanh, the activation function that processes the data on the state and output is tanh in LSTM. $\overrightarrow{C}_t$ is added to the state, and the old cell state C_{t-1} is updated to C_t. Second, the cell updates useful information into cell status and multiply the old cell state C_{t-1} and the output of "forget gate" f_t as the part input of cell, then summing it with the product of "input gate" output i_t and candidate information $\overrightarrow{C}_t$. The result of the calculation is the updated C_t. This is the new candidate and it changes based on how much we decide to update each state. The calculation processes are as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{4}$$

$$\overrightarrow{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \vec{C}_t, \tag{6}$$

Finally, the output value is determined based on the filtered cell state. Firstly, the sigmoid layer determines the output portion of the cell state. Then, the cell state is passed through tanh and multiplied by the output of the sigmoid layer to obtain the result of the cell. The value range of the sigmoid function is [0,1], which is most suitable for controlling the opening and closing of various doors. This part of the calculation processes is shown as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{7}$$

$$h_t = o_t * \tanh(C_t), \tag{8}$$

The key to the LSTM network is the cell state. As shown in Figure 3, the calculation process runs through the horizontal line. It runs directly across the chain with only a small amount of linear interaction and it will be easy to keep the information flowing on it.

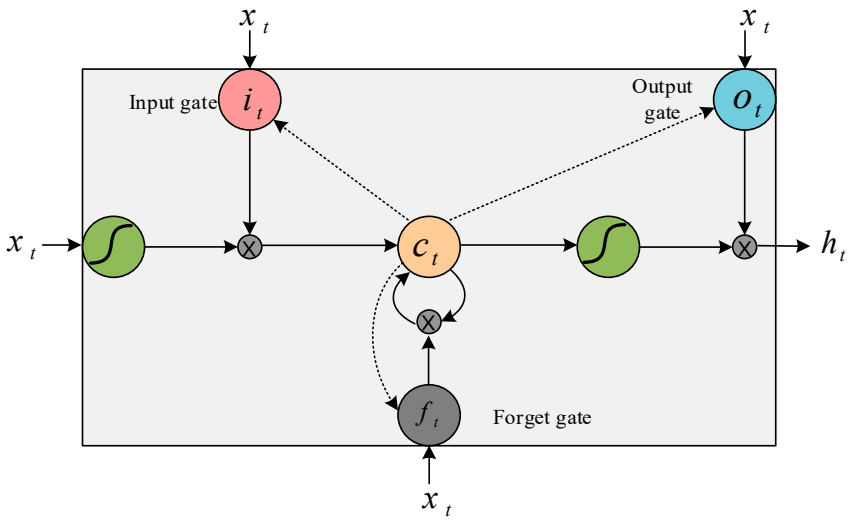


Figure 3. LSTM unit structure.

Therefore, LSTM network can successfully discover sequence relationships and we have added an attention mechanism based on it. CNN network extracts the spatial features to obtain fixed-length feature vectors, and the LSTM network identifies the video contents based on the input feature vectors. We use the framework described in Figure 4, which combines attention-based LSTM with a deep CNN to train spatial-temporal features on video sequences.

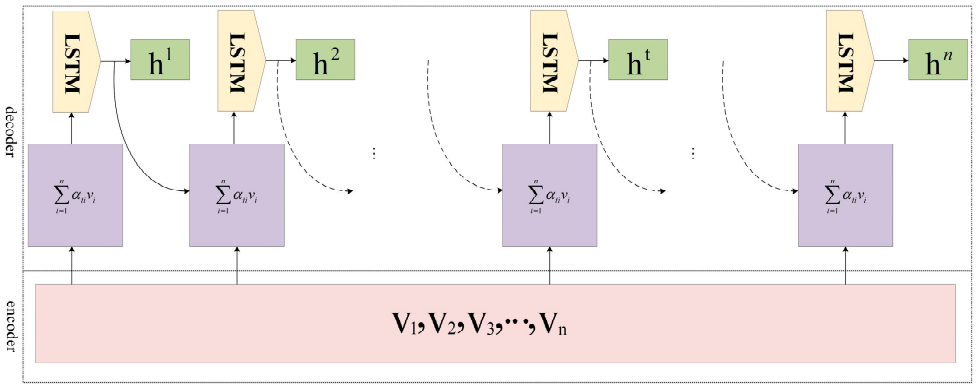


Figure 4. Attention-based LSTM model.

The CNN is used as the encoder and the LSTM network is used as the decoder. In the decoding process, we introduce the attention mechanism and learn the attention weights (α), thus the model pays more attention to the effective area of the whole video [23]. The feature vectors for each frame are weighted and then all video frame sequences (v) are simultaneously used as input $\phi(V)$ to the LSTM network. The input to the attention-based LSTM model is as follows:

$$\phi(V) = \sum_{\mathrm{i}=1}^{n} \alpha_{ti} v_i, \tag{9}$$

The learning of weight α is related to the state of a hidden layer unit on the LSTM network and the feature vector of the current time. The correlation score of α_{ti} is as follows:

$$e_{ti} = \tanh(W \cdot h_{t-1} + U \cdot v_i + b), \tag{10}$$

where h_{t-1} is the output of the hidden unit state at time $t-1$, v_i is the eigenvector of the video frame i, and W, U, b respectively represent the weight matrix to be learned and the offset parameters, the activation function is tanh. Normalization can be obtained as follows:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{\mathrm{n}} \exp(e_{tk})}, \tag{11}$$

where α_{ti} represents the conditional probability ($P_{(a|e)}$) of the video feature vector of the video frame i at time t and the entire video feature vector, furthermore, $\sum_{k=1}^{n} a_{ti} = 1$. The closer the relationship between the frame and whole video feature vector, the bigger the attention weight will become. Then the attention-based LSTM network input at time t is as follows:

$$h_t = f_{rnn}(h_{t-1}, \phi(V)), \tag{12}$$

where f_{rnn} is a unit of LSTM, h_{t-1} is the state of the hidden layer unit at time $t-1$, and $\phi(V)$ is the input at time t after increasing the attention weights.

Although the introduction of the attention mechanism will increase the amount of computation, it can selectively focus on the effective information in the video and reduce the interference of invalid information, thus the performance level of the network model can be significantly improved.

2.3. CNN-LSTM With Attention Networks

Previous studies have shown that both CNN and RNN models can achieve better lip-reading recognition performance alone [24]. We have found that the hybrid network of attention-based CNN-LSTM can further improve performance. The sequence-based attention mechanism can be applied to tasks related to time-series computer vision and assist the model in focusing on some sequence information of the video.

Considering the influence of light, angle and clarity of the input images, we use a better quality of the camera, and the proposed model is trained with RGB images. The part of CNN is improved by using the model based on VGG19 and it does not include the last two fully connected layers (gray parts of Figure 5), and we continue training the model based on pre-training parameters of ImageNet. The structure diagram of VGG19 is shown in Figure 5, and the input of VGG19 is 224 × 244 pixel RGB image. Thus, the output of CNN is 4096 × 10 and the attention mechanism is introduced to the LSTM network to weight keyframes [25]. Thereafter, the network increases two fully connected layers and a SoftMax layer for classification. The calculation process of SoftMax is as follows:

$$f\left(z_j\right) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}}, \tag{13}$$

where z_j represents the output value of the current time j, z_i represents the output value of the time i, and $f(z)$ is the value of the function SoftMax map.

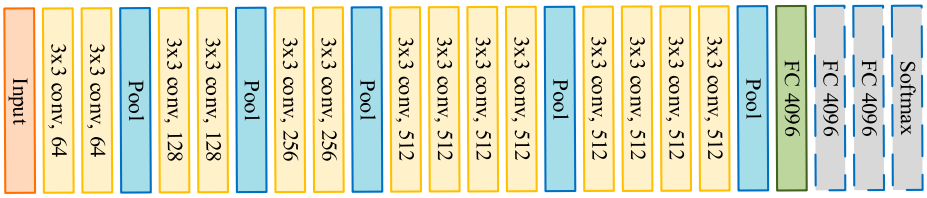


Figure 5. VGG19 network structure.

Obviously, SoftMax converts the output to the probability of each result. The probability value falls within the interval [0,1], and the probability sum is one. This form of results facilitates subsequent calculations.

Finally, the model output is a 10-dimensional vector, and the highest prediction score is obtained as the recognition result according to the SoftMax calculation method.

3. Experimental Dataset and Results

3.1. Dataset

The experimental dataset was performed on the audio-visual database we had created, and 10 independent digital English utterances (numbers from zero to nine) were gathered from six different speakers (three males and three females). Each speaker pronounced each word up to 100 times. The dataset was based on American English pronunciation and the pronunciation of each number was divided into separate video clips. Each independent speaker was not trained in professional pronunciation, and their first language was not always English. Thus, there might be some differences in the lip movements of individual utterances. We collected numerous non-standard samples for the dataset in order to facilitate a more extensive study of the lip-reading recognition system. We collected videos from the frontal perspective of individual speakers who were sitting naturally without any actions. The size of each frame of the original images was 1920 × 1080 resolution, approximately 25 frames per second. In order to accurately locate the beginning and end of each utterance unit, we used audio as an aid to separate each uttered word, each word lasted about 1 s. Then, each isolated word video was further extracted into a fixed length of 10 frames. After processing each video frame, we obtained a fix dimension of ROI at 224 × 224 pixels as standard inputs of CNN model.

3.2. Results and Discussions

In this section, we evaluated our proposed neural network model, and the results were analyzed and compared in our dataset. The out-of-order dataset was randomly divided into 80% training dataset and 20% test dataset. We used pytorch toolkit to carry out the CNN and the attention-based LSTM network. It used a random gradient descent method to train the network in small batches of 50 units with the learning rate of 0.001. The weight of CNN was based on the parameters of the pre-trained model of VGG19, and the weight of attention-based LSTM and the full connection were randomly initialized. The visualization of CNN model was shown in Figure 6.

In order to evaluate the improvement performance influenced by the attention mechanism, we tested and compared the general CNN-RNN architecture. As shown in Figure 7, the proposed network only added one calculation step (gray part) of the attention layer, and the other parts of the two architectures were identical to the initial parameters of the architecture.

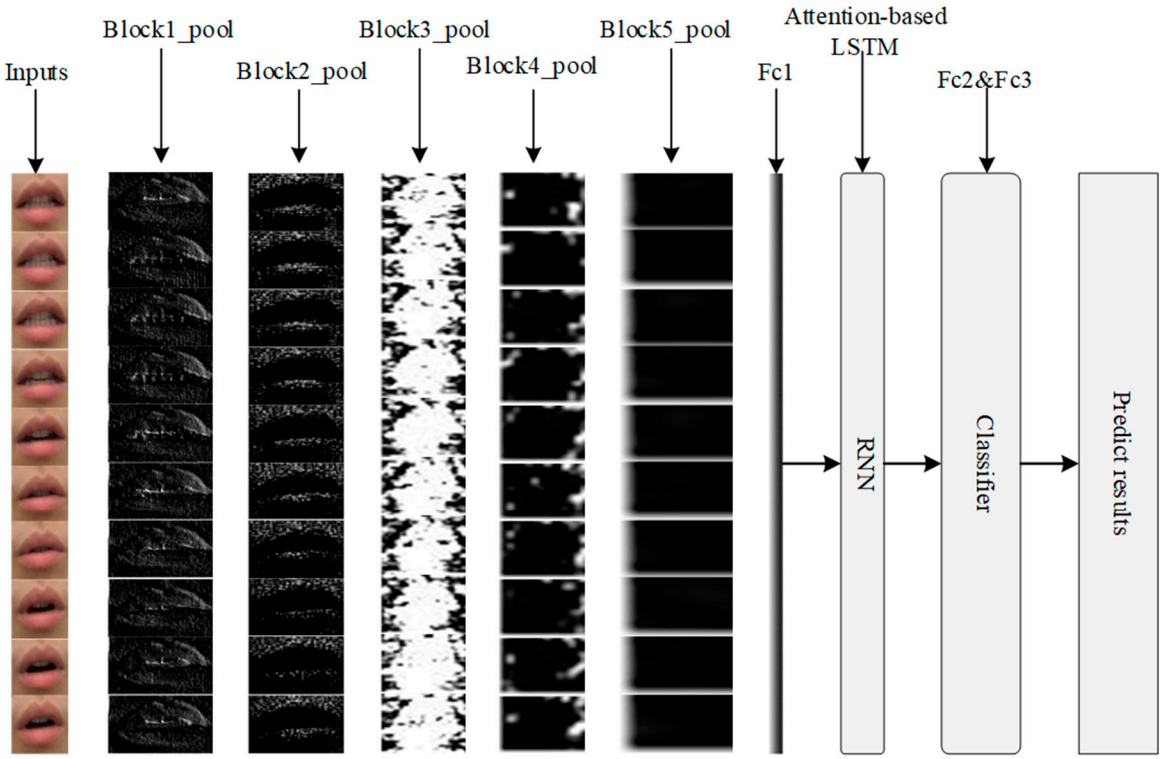


Figure 6. The visualization processes of feature extraction.

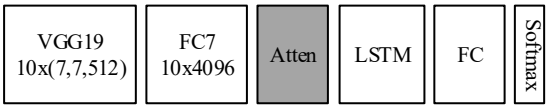
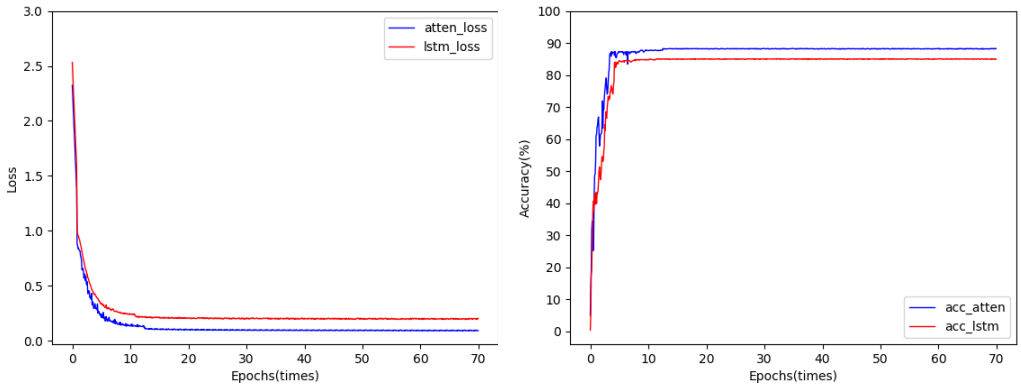


Figure 7. Our proposed architecture and the compared network architecture.

The training dataset and the test dataset were respectively input into two CNN-RNN networks, then we used the same CNN (VGG19) to extract the sequence features of 4096 × 10 and input them into two different RNN networks. The losses, accuracies, and visualizations of the attention mechanism for each period were shown in Figures 8 and 9.



(**a**) (**b**)

Figure 8. Comparison of our proposed attention-based model (blue) with CNN-LSTM model (red). (**a**) Comparison of losses between the two networks and (**b**) comparison of the accuracies between the two networks.

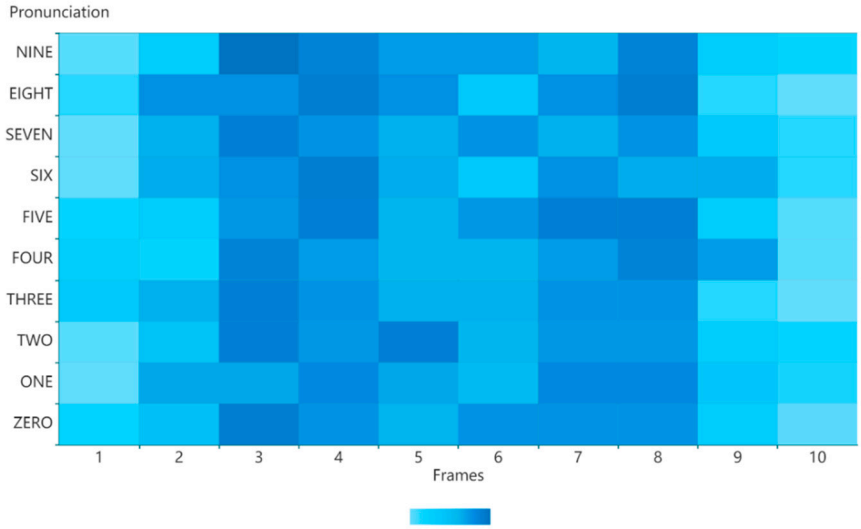


Figure 9. Visualization of the attention mechanism.

In Figure 8, it could be seen that the recognition accuracy of the model with the attention mechanism was higher. When the epochs were around 15 times, the loss tended to be stable and it indicated that the optimal solution had been reached at this time. The accuracy of the proposed network model was 88.2% in the test dataset and 84.9% for the contrastive network (CNN-LSTM).

Figure 9 shows the visualization of the attention mechanism, each line shows the weight of attention at each moment from zero to nine, a deeper color represents a greater weight of attention. It could be inferred that the attention area of each pronunciation was concentrated near the third and seventh frames, because the third, fourth, seventh, and eighth frames contained the main information of the lip motion. These frames were related to the video theme and contained a certain chronological order, which were considered to be important video frames, the model assigned a large amount of attention weight. This attention distribution showed that the attention mechanism optimized the proportion of keyframe and achieved the requirement of allocating more information processing resources to important parts. Therefore, our research used an architecture which was a fusion of CNN and attention-based LSTM.

Obviously, the performance of our proposed architecture was more powerful. We tested the English words in ten independent utterances in two models. The experimental results were shown in Figure 10. The results show that the proposed model had higher accuracy in all independent digital word utterances and it was stronger than the general CNN-LSTM model. Furthermore, according to the analysis of the recognition results for each individual utterance of English words, the identification of "two" was the easiest, "zero" was the most difficult to identify. Complex pronunciation was difficult to recognize because of errors in individual pronunciation. The best-recognized utterance was "two" because the speakers' movements were consistent without regional differences, and the worst identified utterance was "zero" because the lip movements of speakers were complicated and a part of pronunciation required tongue to control syllables. Since the experimental data was not using the first language of all speakers, the experimental result would be adversely affected. It could be inferred that the accuracy of lip-reading recognition result would be higher if standard announcer's pronunciation videos were used. In addition, it was difficult to put it into practical applications. Therefore, the dataset in this paper was closer to the actual application and it had high academic research value. As a whole, the proposed model effectively improved recognition accuracy. It could be concluded that the proposed model was stronger than the general CNN-LSTM structure in the performance of these ten independent digital pronunciations.

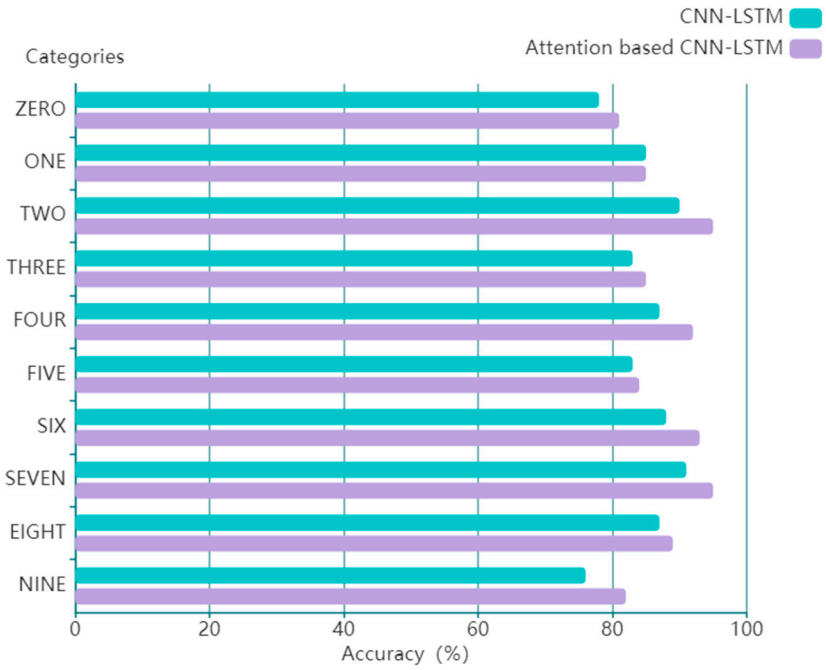


Figure 10. Comparisons of recognition accuracy between two networks.

4. Conclusions

In this paper, hybrid neural network architecture of CNN and attention-based LSTM is proposed for lip-reading recognition systems. Firstly, CNN (VGG19) extracted visual features from the mouth ROI. Then, we used the attention-based LSTM to learn the sequence weights and sequence information between the frame-level features. Finally, the classification was achieved by using two fully connected layers and a SoftMax layer. The experimental dataset was built by us independently and it consisted of three males and three females. American English pronunciation of numbers from zero to nine, and each digital utterance were divided into independent video clips, each independent speaker was not trained in professional pronunciation. The experimental results show that compared with the general CNN-RNN model, the proposed architecture can effectively predict words from the sequence of lip region images on our own dataset, and the accuracy of the proposed model is 88.2% in the test dataset which is 3.3% higher than the general CNN-RNN. In future research, we will train the lip-reading recognition model on datasets of real-time broadcast videos, including video samples from news broadcasts and real-world environments to explore our proposed approach for speaker-independent video speech recognition system.

Author Contributions: Data curation, Y.L. and H.L.; Formal analysis, Y.L. and H.L.; Methodology, Y.L. and H.L.; Project administration, Y.L.; Resources, Y.L.; Supervision, Y.L.; Validation, H.L.; Visualization, H.L.; Writing—original draft, H.L.; Writing—review & editing, Y.L. and H.L.

Funding: This research was supported by the National Natural Science Foundation of China (61571013), by the Beijing Natural Science Foundation of China (4143061), by the Science and Technology Development Program of Beijing Municipal Education Commission (KM201710009003) and by the Great Wall Scholar Reserved Talent Program of North China University of Technology (NCUT2017XN018013).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jaimes, A.; Sebe, N. Multimodal human–computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134. [CrossRef]
2. Loomis, J.M.; Blascovich, J.J.; Beall, A.C. Immersive virtual environment technology as a basic research tool in psychology. *Behav. Res. Methods Instrum. Comput.* **1999**, *31*, 557–564. [CrossRef] [PubMed]

3. Hassanat, A.B. Visual passwords using automatic lip reading. *arXiv*, 2014; arXiv:1409.0924.
4. Thanda, A.; Venkatesan, S.M. Multi-task learning of deep neural networks for audio visual automatic speech recognition. *arXiv*, 2017; arXiv:1701.02477.
5. Biswas, A.; Sahu, P.K.; Chandra, M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *Int. J. Speech Technol.* **2016**, *19*, 159–171. [CrossRef]
6. Scanlon, P.; Reilly, R. Feature analysis for automatic speechreading. In Proceedings of the 2001 IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564), Cannes, France, 3–5 October 2001; pp. 625–630.
7. Matthews, I.; Potamianos, G.; Neti, C.; Luettin, J. A comparison of model and transform-based visual features for audio-visual LVCSR. In Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2001, Tokyo, Japan, 22–25 August 2001; pp. 825–828.
8. Aleksic, P.S.; Katsaggelos, A.K. Comparison of low-and high-level visual features for audio-visual continuous automatic speech recognition. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; p. V-917.
9. Chitu, A.G.; Driel, K.; Rothkrantz, L.J. Automatic lip reading in the Dutch language using active appearance models on high speed recordings. In Proceedings of the International Conference on Text, Speech and Dialogue, Brno, Czech Republic, 6–10 September 2010; pp. 259–266.
10. Luettin, J.; Thacker, N.A.; Beet, S.W. Visual speech recognition using active shape models and hidden Markov models. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; pp. 817–820.
11. Shaikh, A.A.; Kumar, D.K.; Yau, W.C.; Azemin, M.C.; Gubbi, J. Lip reading using optical flow and support vector machines. In Proceedings of the 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; pp. 327–330.
12. Puviarasan, N.; Palanivel, S. Lip reading of hearing impaired persons using HMM. *Expert Syst. Appl.* **2011**, *38*, 4477–4481. [CrossRef]
13. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
14. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. In Proceedings of the 9th International Conference on Artificial Neural Networks: ICANN'99, Edinburgh, UK, 7–10 September 1999.
15. Wang, Y.; Huang, M.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 606–615.
16. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014; arXiv:1409.1556.
17. Matthews, I.; Cootes, T.F.; Bangham, J.A.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213. [CrossRef]
18. Fan, X.; Zhang, F.; Wang, H.; Lu, X. The system of face detection based on OpenCV. In Proceedings of the 2012 24th Chinese Control and Decision Conference (CCDC), Taiyuan, China, 23–25 May 2012; pp. 648–651.
19. King, D.E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
20. Martins, A.; Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1614–1623.
21. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
22. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
23. Cho, K.; Courville, A.; Bengio, Y. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. Multimed.* **2015**, *17*, 1875–1886. [CrossRef]

24. Zhang, Y.; Pezeshki, M.; Brakel, P.; Zhang, S.; Bengio, C.L.Y.; Courville, A. Towards end-to-end speech recognition with deep convolutional neural networks. *arXiv*, 2017; arXiv:1701.02720.
25. Graves, A. Generating sequences with recurrent neural networks. *arXiv*, 2013; arXiv:1308.0850.