The Research of Lip Reading Based on STCNN and ConvLSTM

Yijie Zhu
School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100000, China
E-mail: zhuyijie@bjtu.edu.cn

Abstract. Aiming at the problems in temporal model during the research of lip reading, a deep learning model is proposed based on spatiotemporal convolutional neural networks (STCNN) and Convolutional Long Short-Term Memory (ConvLSTM). Firstly, STCNN is used to learn the features of the extracted lip image, and then the learned features are sent to ConvLSTM to process the time series data, which is classified by softmax, and finally the CTC loss function is used to optimize the results. Using GRID data set for training, comparing with experiments, it is found that the recognition accuracy of this model achieves 95.0% at the word level. Experiments show that the model can improve the accuracy of lip reading.

1. Introduction

Lip reading is a complex technology which integrates computer vision and natural language processing. The system can recognize the face continuously from the video, extract the sequence of the features of the mouth shape, and then input the sequence into the recognition model, recognize the corresponding pronunciation of the speakers’ mouth shape, and then calculate the most likely expression statement. This technology has a wide range of application prospects in the fields of espionage, disabled equipment and so on.

Lip reading includes lip detection and location, feature extraction, language model and target recognition module. It is an interdisciplinary field of speech signal processing and image processing. Most of the researches focus on lip reading, lip feature extraction and deep learning network training. In the aspect of using deep learning network for training, there are many kinds of deep learning network models, such as LIPNET from Y M Assael[1].

In this paper, a deep learning method based on spatiotemporal convolution network (STCNN) and convolutional Long Short-Term Memory (ConvLSTM) is proposed, which uses the influence of temporal and spacial information on cyclic training to achieve lip reading of English short sentences.

2. LIPNET

The LIPNET model is mainly composed of the following parts: (1) Use STCNN to learn lip features, (2) Use Bidirection Gated Recurrent Unit (Bi-GRU) model to extract temporal features of lip movement, (3) Use connectionist temporal classification (CTC) loss function to adjust model parameters. As shown in figure 1.
2.1. Spatiotemporal convolutions

LIPNET uses spatiotemporal convolution network (STCNN). Convolution neural network (CNN) includes stacked convolution which can be operated on a single image in space[2], while spatiotemporal convolution neural network introduces time dimension to process video data in temporal and spacial dimension[3][4]. Formally, the value of position \((x, y, z)\) on the \(j\)-th characteristic graph of layer \(i\) is given by the following formula: \(k_t\) is the size of 3D core along the time dimension, \(w\) is the weight, and \(x\) is the input.

\[
[\text{stconv}(x, w)]_{i,j} = \sum_{c=1}^{C} \sum_{f=1}^{k_t} \sum_{i=1}^{k_i} \sum_{j=1}^{k_j} w_{c,f} x_{c,f,i,j} \quad (1)
\]

2.2. Bidirection Gated Recurrent Unit

Bidirection Gated Recurrent Unit (Bi-GRU) structure used in the circulation layer to process temporal sequences and it is a type of recurrent neural network (RNN). Bi-GRU is based on Gated Recurrent Unit (GRU)[5]. GRU is a variant of LSTM structure[6]. The structure is showed in figure 2.

There is a current input \(x_t\) and the hidden state \(h_{t-1}\) passed down from the previous node, which contains information about the previous node.

GRU obtains the two gating states by the last transmitted state \(h\) and the input \(x\) of the current node. As shown in (2)(3), where \(r_t\) is the reset gate for control reset, \(z_t\) is the update gate for control update, and \(\sigma\) is the sigmoid function.

\[
r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (2)
\]

\[
z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (3)
\]
After getting the gating signal, the reset gating is used to get the reset data $\hat{h}_{t-1} r_t$, and then it is spliced with the input $x_t$. Then, a tanh activation function is used to zoom the data to the range of (-1,1). That is, $\hat{h}_t$ is obtained as following formula (4). Here $\hat{h}_t$ mainly contains the current input data $x_t$. Adding $\hat{h}_t$ to the current hidden state pertinently is equivalent to memorizing the state of the current moment.

$$\hat{h}_t = \text{tanh}(W \cdot [x_t, h_{t-1}])$$ \hspace{1cm} (4)

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \hat{h}_t$$ \hspace{1cm} (5)

The range of gating signal $z_t$ is (0,1). The closer the gating signal is to 1, the more data will be memorized; and the closer to 0, the more data will be forgotten.

Bi-GRU adds reverse calculation on the basis of forward calculation, which is better for time series information processing[7]. Structure is showed in figure 3.

![Figure 3. Gated Recurrent Unit](image)

2.3. Connectionist Temporal Classification

LIPNET model uses Connectionist Temporal Classification (CTC) loss function[8] to optimize training results. The CTC loss function allows the model not to be rigid to align the input and target outputs in the training data. CTC loss function is trained based on sequence, so it allows neural network to predict in any period. Given a model that outputs a discrete distributed sequence on a token, which is enhanced by a special blank token, CTC calculates the probability of the sequence by marginalizing all sequences defined as equivalent.

$$\tilde{V} = V \cup \{\}_V$$ \hspace{1cm} (6)

The space symbol indicates the blank of CTC. Define the function $B: \tilde{V} \rightarrow V^*$, give a string on $\tilde{V}$, delete adjacent duplicate characters and remove the blank token. For a tag sequence $y \in V^*$, CTC defines:

$$p(y | x) = \sum_{u \in B^{-1}(y) \cap\{|B/u|=T\}} p(u_1, ..., u_T | x)$$ (7)

In the formula, $T$ is the number of time steps in the sequence model. If $T=3$, there are 5 possibilities: $p(iis)$, $p(iis)$, $p(\_is)$, $p(\_s)$, $p(is\_)$.
3. Model improvement

For the original LIPNET model, we propose a model based on the original model. The main improvement is to replace the Bi-GRU cycle layer of the original model with convolutional LSTM network[9].

Convolutional LSTM network (ConvLSTM) introduces the idea of convolution on the basis of LSTM. It uses convolution to extract features, instead of the original full connection method. The traditional LSTM is 1D tensor, and the input of convLSTM is 3D tensor. ConvLSTM is more effective for image feature extraction. The basic structure of ConvLSTM is shown in figure 4.

![Figure 4. Convolutional LSTM](image)

ConvLSTM is a variant of LSTM, which has the basic characteristics of LSTM. It uses three gates to control cell state, which are input gate, output gate and forgetting gate.

The first step in LSTM is to determine what information the cell state needs to discard. This part of the operation is achieved by forgetting door. The forgetting gate consists of a sigmoid function. Sigmoid unit processes the input $H_{t-1}$ and $x_t$, then outputs $f_t$, the value of $f_t$ controls whether the information in $c_{t-1}$ is retained or not. 0 is not reserved and 1 is completely reserved. Different from the original LSTM, convLSTM uses convolution instead of simple corresponding multiplication to calculate the weights of the input $h_{t-1}$ and $x_t$.

$$i_t = \sigma(w_{si} * x_t + w_{hi} * h_{t-1} + w_{ci} * c_{t-1} + b_i)$$

The input gate is used to determine which information needs to be updated. In the first step, $h_{t-1}$ and $x_t$ are processed with sigmoid unit elements to obtain $i_t$, $i_t$ is a vector with a value between 0 and 1, the value of $i_t$ represents the weight of information in $C_t$, 0 is unimportant, and 1 is important. Then, $h_{t-1}$ and $x_t$ are processed with tanh layer to obtain new cell information $\hat{C}_t$, where tanh represents hyperbolic tangent function and acts on every element in the vector.

$$f_t = \sigma(w_{sf} * x_t + w_{hf} * h_{t-1} + w_{cf} * c_{t-1} + b_f)$$

$$\hat{C}_t = \tanh(w_{sc} * x_t + w_{hc} * h_{t-1} + b_i)$$

The Hadamard product of $f_t$ and $C_{t-1}$ and the Hadamard product of $i_t$ and $g_t$ are summed to obtain new cell information $C_t$.

$$C_t = f_t \circ C_{t-1} + i_t \circ \hat{C}_t$$

The output gate is used to control the output of $c_t$, expressed as $o_t$. The sigmoid unit of the output gate will judge the input $h_{t-1}$ and the $x_t$, then process the new cell state $c_t$ through the tanh function, and then calculate the Hadamard product with $o_t$ to get the final output result.

$$o_t = \sigma(w_{so} * x_t + w_{ho} * h_{t-1} + w_{co} * c_{t-1} + b_o)$$
\[ h_t = o_t \odot \tanh(C_t) \]  

Compared with LSTM and GRU models, convLSTM can not only establish temporal model, but also use CNN characteristics to describe local spatial features. ConvLSTM is better than LSTM in obtaining spatiotemporal relationship and solving lip reading problem better.

The improved model is shown in figure 5. STCNN consists of three layers of convolution layer and maximum pooling layer, and then uses bidirectional ConvLSTM to process the image sequence. Each output is processed by a linear layer and softmax classifier. Finally, This model is trained and filtered by CTC loss function.

![Figure 5. Improved model](image)

4. Evaluation

4.1. Data set
The research uses GRID data set[10], including 32746 English sentences. The standard form of each sentence is \(<\text{verb}> + \text{<color>} + \text{<preparation>} + \text{<digit>} + \text{<letter>} + \text{<advertiser}>\). A total of 34 people participated in the recording, including 18 males and 16 females, each recording 1000 sentences, and some sentences were missing. Each sentence is 3 seconds, the frame number is 25F/s. The research choose 160*80 pixels set.

4.2. Setup
We employ a split (unseen speakers; not previously used in the literature) holding out the data of two male speakers and two female speakers for evaluation, 3971 videos in total. The remaining 28775 videos are used for training. Another split (overlapped speakers) includes 255 random sentences from each speaker. 24408 videos are used for training and 8415 videos are used for validation. The videos were processed with the Dlib face detector[11], and the iBug face landmark predictor with 68 landmarks[12]. Using these landmarks, an affine transformation is applied to extract a mouth-centred crop of size 128×64 pixels per frame. The RGB channel is standardised to have zero mean and unit variance. The neural network model is built by the keras Library of Python language and based on tensorflow back end. The GPU used in the experiment is GTX Titan X.

Two control experiments are set up: LIPNET and improved model.

4.3. Result
To evaluate the performance of the two models, character error rate (CER) and word error rate (WER) are computed as standard metrics for the performance of ASR models. In order to make the identified word (character) sequence consistent with the standard word (character) sequence, it is necessary to replace, delete or insert some words (characters). The total number of these inserted, replaced or deleted words divided by the percentage of the total number of words (characters) in the standard sequence is WER (CER).
Table 1. Results

| Scenario                | CER  | WER  |
|-------------------------|------|------|
| Unseen speakers (Origin)| 6.7% | 13.6%|
| Overlapped speakers (Origin)| 2.0% | 5.6% |
| Unseen speakers (Ours)  | 6.4% | 12.9%|
| Overlapped speakers (Ours)| 1.9% | 4.5% |

The model improves the character error rate by 0.3%, the word error rate by 0.7%, on unseen speakers set, and the character error rate by 0.1%, the word error rate by 1.1% on overlapped speakers set. In ConvLSTM, time series data is image. Convolution operation is added on the basis of LSTM, which is effective for image feature extraction. The experiment shows that the recognition accuracy of the method with ConvLSTM is improved.

5. Conclusion
Through the researches on the existing deep learning and lip reading, it is found that there is a problem of insufficient utilization of spatial information by circular network. To solve this problem, this paper proposes a lip reading method based on STCNN and Bi-ConvLSTM, which can effectively learn the spatiotemporal features of lips, and capture the spatiotemporal information and context information through Bi-ConvLSTM.

The experimental result shows that the lip reading method combined with STCNN and Bi-ConvLSTM can improve the accuracy of lip reading.

This experiment still needs to be improved. GRID data set is inflexible, which makes it inconvenient for corpus expansion. Because ConvLSTM uses convolution method, the amount of computation increases, which leads to long training time. These problems will be improved in future research.

Acknowledgments
Thanks to Undergraduate Innovation and Entrepreneurship Training Program of Beijing Jiaotong University (1901wh201) for the high performance server.

References
[1] Assael, Y.M., Shillingford, B., Whiteson, S., Freitas, N. de. (2017). LipNet: End-to-End Sentence-level Lipreading. ArXiv Preprint ArXiv:1611.01599.
[2] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of The ACM, 60(6), 84–90.
[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks. In CVPR ’14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus. pp. 1725–1732.
[4] Ji, S., Xu, W., Yang, M., Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1), 221–231.
[5] Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv Preprint ArXiv:1412.3555.
[6] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[7] Graves, A., Schmidhuber, J. (2005). 2005 Special Issue: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5), 602–610.
[8] Graves, A., Fernández, S., Gomez, F., Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning. pp. Orlando. 369–376.

[9] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., Woo, W. (2015). Convolutional LSTM Network: a machine learning approach for precipitation nowcasting. In NIPS’15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. Montreal. pp. 802–810.

[10] Cooke, M., Barker, J., Cunningham, S., Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America, 120(5), 2421–2424.

[11] King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research, 10(60), 1755–1758.

[12] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M. (2016). 300 Faces In-The-Wild Challenge. Image and Vision Computing, 47, 3–18.