Lip-Corrector: Application of BERT-based model in sentence-level lipreading

Haoran Zhao¹, Bowen Zhang¹, Zhanhang Yin¹

¹School of Artificial Intelligence, Hebei University of Technology, Tianjin, 300131, China

*Corresponding author’s e-mail: 184965@stu.hebut.edu.cn

Abstract. The current field of lipreading is limited to the processing of visual signal and the optimization of sequence models, but the sentence text is ignored. Aiming at this problem, we proposed a lipreading method combined with natural language processing (NLP) technology, Lip-Corrector, which applies the BERT model in this paper. The front end of the model uses 3D+2D convolutional neural network (CNN) to extract lip information, the middle end uses the Transformer-based Seq2seq sequence model to make sentence-level predictions, and the back end uses a sentence correction method based on the BERT model, which connects to the middle after pre-training on the self-made dataset. Experiments on the two largest sentence-level lipreading datasets of LRS2 and LRS3 show that the performance of this model surpasses all the baselines, which proves that lipreading methods combined with NLP technology will get better results.

1. Introduction

Lipreading refers to predicting the content of the speaker by obtaining the lips motion in the video, which is very difficult for humans to complete. On the one hand, different people have different speaking habits, their lip movement features are different from each other, on the other hand, the lip movement is not large, it is difficult to distinguish the lip shapes of many pronunciations. Recently, this problem is gradually solved by computers. Models such as CTC [4] and Seq2seq[5] have been applied to sentence-level lipreading, which have achieved good results. Many models have already surpassed the performance of humans.

The current mainstream lipreading methods focus on image processing and sequence model optimization in the field of computer vision(CV). However, lipreading involves two parts: visual signal processing and text processing. So, researchers can not only focus on the CV part, should also use natural language processing (NLP) technology to process text. As far as the current research is concerned, researchers have not paid enough attention to the processing of lipreading text. Through the observation of a large number of sentence-level lipreading results, we found that many results have problems that can not be eliminated in context. These problems are mainly manifested in the fact that some words predicted incorrectly confuses the meaning of the whole sentence.

This paper proposes a BERT-based lipreading model Lip-Corrector. As far as we know, this is the first model that combines NLP models with traditional lipreading methods. Lip-Corrector consists of three parts, like the modern lip language recognition model, its front-end extracts lip features through a convolutional neural network(CNN), its middle-end model is used for sentence-level sequence prediction and its back-end uses a sentence correction model based on the BERT model. Through pre-training on the self-made sentence correction datasets, BERT based model connects with the middle
Experiments on the two largest sentence-level lipreading datasets of LRS2 and LRS3 show that the performance of this model surpasses all the baselines.

2. Related work

2.1. Sentence-level lipreading
Sentence-level lipreading is one of the current research hotspots in deep learning. Assael et al.[1] used the Bi-LSTM+CTC method to achieve sentence-level lipreading in the GRID dataset in 2016. However, the GRID data set sentence does not have natural language features. The subsequent LRS2-BBC, LRS3-TED, etc. belong to the formal sentence-level recognition. In 2017, Chung et al.[2] applied the Attention mechanism to LSTM and achieved good results. In 2018, Afouras et al. [3] combined the CTC model with the Transformer to achieve the latest results.

2.2. Sentence correction
Due to the negligence or the insufficient ability to master the language of the writer, writing errors are very common. Common sentence errors are mainly divided into two types[6]: (1) real-word errors, which means words are misspelled but their form is in the correct word dictionary, and (2) non-word errors, which means the incorrect word form is not in the correct word dictionary. For the lipreading text problem, we are more inclined to the latter. Since the lip recognition model will produce classification errors, it may be predicted as a wrong word at a specific location. The word form is in the correct dictionary, but it is not the correct word of the sentence. In this article, we improved the BERT model[7] and applied it to lipreading. The BERT model was proposed in 2018 and subsequently attracted widespread attention in the NLP field. It has shown excellent results in sentence translation, spelling verification and other NLP problems.

3. Lip-Corrector

3.1. Model structure
The Lip-Corrector model consists of three parts, as shown in Figure 1, the front-end uses 3D+2D convolution to extract image features, the middle-end uses a Transformer-based Seq2seq model and the back-end is connected to pre-trained BERT.

![Figure 1. Lip-Corrector model structure diagram.](image)

3.2. 3D+2D Convolutions
Due to the temporality of lipreading and the 2D convolutional network cannot handle the temporal structure, Stafylakis and Tzimiropoulos [12] using 3D and 2D convolutional neural network to process the features of time information. After 2D CNN operation, the time information of the input data will be lost, but the 3D CNN can stack the time information of consecutive frames into a multi-dimensional cube[15].

As shown in Figure 2, the input data is first preprocessed into grayscale sequence images, and then the data is input to the 3D CNN and 2D ResNet. Among them, 3D CNN has one layers and ResNet has 34 layers to extract motion features of lip.
3.3. Transformer-based Sequence-to-sequence

The transformer-based sequence-to-sequence (TM-Seq2seq) model was first proposed by Afouras et al. [14]. In this model, a separate attention head is used to participate in video embedding. In each decoder layer, the resulting video contexts are concatenated together on the channel and propagated forward. The character probability generated by the model matches the label and is trained through cross entropy loss.

Both the self-attention and encoder-decoder attention layers of the model are used in the same multi-head attention block. As described by Vaswani et al. [8], a query (Q), a key (K), and a value (V) tensor are received as input, and h context vectors are generated. Each attention head $i$:

$$\text{Att}_i(Q, K, V) = \text{softmax} \left( \frac{\left( W_i^Q Q^T \right)^T \left( W_i^K K^T \right)}{d_k^{1/2}} \right) \left( W_i^V V^T \right)^T$$

Embed the video with a separate attention header. In each decoder layer, the resulting videos are connected together in the channel dimension and transmitted to the input block. All attention mechanisms use the output of the previous decoding layer as a query. The character probability generated by the decoder directly matches the true value label and undergoes cross-entropy loss training.

3.4. BERT-based model

Devlin et al. [7] used two unsupervised methods to pretrain the BERT model when they proposed BERT. In this article we have adopted one of them, Masked LM (MLM) method, randomly erase one or several words in a sentence and predict the erased words based on the remaining words. For words that are erased in the original sentence, a special symbol MASK is used in 80% of cases, an arbitrary word is used in 10% of cases, and the original word is kept unchanged in the remaining 10%. After that, the model does not know whether each corresponding vocabulary is the correct vocabulary, which forces the model to rely more on contextual information to predict vocabulary, and gives the model a certain error correction ability.

Transformer is the core structure of the BERT model. The BERT model adds several key operations on the basis of Multi-head Self-Attention: (1) Residual connection: The input and output of the module are directly added as the final output. Such changes can make it easier to modify the model, that is, modify the output by modifying the weight of the input (2) Layer Normalization: normalize a layer of neural network nodes with 0 mean and 1 variance (3) Linear transformation: the enhanced semantic vector of each word is performed two more linear transformations to enhance the expressive ability of the entire model. The transformed vector keeps the same length as the original vector.
4. Datasets

4.1. Lipreading datasets

**LRS2-BBC** contains thousands of spoken sentences from BBC television. The maximum length of each sentence is 100 characters. The statistics of the data set are given. The dataset in testing corresponds to partial sentences and multiple sentences, but the training set contains only one complete sentence. There is some overlap between pre-training and training sets. Although there may be some label noise in the pre-training and training sets, the test set has been additionally verified; therefore, as far as we know, there are no errors in the test set.

**LRS3-TED** includes over 400 hours of videos extracted from 5,594 TED and TEDx English conversations downloaded from YouTube. These videos are provided in the form of .mp4 files with a resolution of 224×224 and a frame rate of 25 fps. Encoding uses the h264 codec. The audio track is provided in a single-channel 16-bit 16kHz format, and the corresponding text record and the alignment boundary of each word are contained in a plain text file. The data set is divided into three groups: pre-training, training and testing.

4.2. Sentence correction datasets

We perform noise processing on sentences to generate sentence data that corrects Sentence errors. We use one million sentences in the billion-character benchmark[9] dataset as the original visual corpus, and use Random noise strategy[10], which involves sperrute, delete, insert and replace operations to add errors in each sentence to obtain a dataset containing about 20% noise. Figure 3 shows some examples of the dataset processing.

![Figure 3. Sentence correction dataset samples.](image)

5. Evaluations

5.1. BERT model evaluation

In the experiment of sentence correction, we chose SC-LSTM and CHAR-LSTM-LSTM to compare with BERT. The experimental results on the self-made data set show that after 300 rounds of training, the BERT model performs best under the two evaluation criteria of Word-Level Accuracy and Correction Rate.

| Methods                  | Accuracy  |
|--------------------------|-----------|
| SC-LSTM[10]              | 95.6/88.5 |
| CHAR-LSTM-LSTM[13]       | 96.0/89.1 |
| BERT[7]                  | **96.9/93.3** |
5.2. Performance Evaluation

Compared with the basic lipreading model TM-Seq2seq, we connect SC-LSTM, SHAR-LSTM-LSTM and BERT with TM-Seq2seq respectively. The results are shown in Table 2. The TM-Seq2seq combined with the NLP model performed better than the baselines and other lipreading methods. The model combined with BERT has the most outstanding performance, and is the best on the two datasets of LRS2-BBC and LRS3-TED.

| Methods                        | LRS2-BBC | LRS3-TED |
|--------------------------------|----------|----------|
| WAS[11]                        | 70.4%    | -        |
| Bi-LSTM+Attention-CTC[12]      | 57.9%    | -        |
| TM-Seq2seq[14]                 | 49.8%    | 59.9%    |
| TM-Seq2seq+SC-LSTM             | 47.9%    | 59.0%    |
| TM-Seq2seq+CHAR-LSTM-LSTM      | 47.6%    | 58.6%    |
| TM-Seq2seq+BERT(ours)          | 47.1%    | 58.2%    |

![Figure 4. Sentence correction dataset samples.](image)

6. Conclusions

We proposed a BERT-based method for sentence-level lipreading, Lip-Corrector, applying the BERT model to the field of lipreading in this paper. We conducted experiments on two large public sentence-level lipreading datasets LRS2 and LRS3, the lowest sentence-level word error rate (WER) reached 47.1% and 58.2%, surpassed current baselines. We finally proved that lipreading combined with NLP technology will get better results.

References

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, “LipNet: End-to-end sentence-level lipreading,” 2016, arXiv:1611.01599.
[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3444–3453.
[3] T. Afouras, J. Son Chung, and A. Zisserman, “Deep lip reading: A comparison of models and an online application,” 2018, arXiv:1806.06053.
[4] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the International Conference on Machine Learning, pages 369–376. ACM, 2006.
[5] K. Cho, B. Van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In EMNLP, 2014.

[6] Hládek, D., Staš, J., & Pleva, M. (2020). Survey of automatic spelling correction. Electronics (Switzerland), 9(10), 1–29. https://doi.org/10.3390/electronics9101670

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In Advances in Neural Information Processing Systems, 2017.

[9] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2014). One billion word benchmark for measuring progress in statistical language modeling. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2635–2639.

[10] Sakaguchi, K., Duh, K., Post, M., & Van Durme, B. (2017). Robust word recognition via semi-character recurrent neural network. 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 3281–3287.

[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[12] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, “Audio-visual speech recognition with a hybrid CTC/attention architecture,” in Proc. IEEE Spoken Lang. Technol. Workshop (SLT), Dec. 2018, pp. 513–520.

[13] Li, H., Wang, Y., Liu, X., Sheng, Z., & Wei, S. (2018). Spelling error correction using a nested RNN model and pseudo training data. ArXiv.

[14] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep Audio-visual Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–13. https://doi.org/10.1109/TPAMI.2018.2889052

[15] Hao, M., Mamut, M., Yadikar, N., Aysa, A., & Ubul, K. (2020). A Survey of Research on Lipreading Technology. IEEE Access, 8, 204518–204544. https://doi.org/10.1109/access.2020.3036865