Lip Reading using Local-Adjacent Feature Extractor and Multi-Level Feature Fusion

Chengwei ZHANG, Haotian ZHAO
School of Artificial Intelligence, Hebei University of Technology, Tianjin 300132, China
184948@stu.hebut.edu.cn

Abstract. Lip reading is a widely used technology, which aims to infer text content from visual information. To represent lip information more efficiently and reduce the network parameters, most networks will first extract features from lip images and then classify the features. In recent studies, most researchers adopt convolutional networks to extract information from pixels which contain a lot of useless information, limiting the improvement of model accuracy. In this paper, we designed a graph structures and a lip segmentation network to effectively represent changes in the shape of the lips in adjacent frames and the ROI in local frame and propose two feature extractors, named U-net-based local feature extractor and graph-based adjacent feature extractor. We proposed a very challenging dataset to simulate extreme environments, including highly variable face properties, light intensity and so on. Finally, we designed several different levels of feature fusion methods. The experimental results on the proposed challenging dataset show that the model can effectively extract the useful information from content irrelevant information very well. The accuracy of our proposed model is 9.1% higher than that of baseline. This result shows that our proposed model can better adapt to the application of the wild environment.

1. Introduction
With the development of science and technology and the improvement of hardware manufacturing level, deep learning based artificial intelligence technology has attracted more and more attention from researchers. The field of deep learning includes many sub-fields, such as computer vision, natural language processing, etc. Lip reading combining computer vision and natural language processing has become a research hotspot. Lip reading has a wide range of application scenarios, such as in-body detection based on lip features, communication assistance for hearing impaired people, traffic camera voice recovery, etc. Lip reading faces many difficulties, such as the speaker’s identity, facial posture, and other external environment such as lighting conditions. These factors will increase the difficulty of recognition task.

At present, there are many solutions to lip reading. Most of them are based on end-to-end lip reading models, which separate the network into feature extractor front-end and feature classifier back-end. The front-end feature extractor extracts the visual features of lips from the original image of lips, and other variants derived from it, such as deformation flow [1] and diff images [2] that depict pixel changes of adjacent images. The better the performance of the feature extraction network, the higher the accuracy of classification. However, most of these feature extraction methods do not consider the influence of light intensity, light angle, and speaker's identity information on the task of lip reading. The ability of traditional convolutional network based on image pixels to extract
information about changes in lip sequences is also very limited, which makes the accuracy of lip reading in a bottleneck period.

In this paper, we will focus on how to extract the lip feature effectively from the irrelevant information and how to fusion the feature extracted from the graph structure and pixel images. Firstly, we designed a graph structures and a lip segmentation network to effectively represent changes in the shape of the lips in adjacent frames and the region of interest in local frame. We propose two feature extractors, named U-net-based local feature extractor and graph-based adjacent feature extractor to extract the useful information from lip images with irrelevant factors. We proposed a very challenging dataset to simulate extreme environments, including highly variables face properties, light intensity and so on. Finally, we designed several different levels of feature fusion methods such as feature-level, score-level, and decision-level. By using the proposed novel extractors and feature fusion methods, the lip reading model can achieve a better performance when the speaker has distinctive facial features or is exposed to extreme light conditions.

2. Architecture

2.1. Local feature extractor

In order to effectively extract visual information from lips images with complex environmental changes and different facial attributes, we designed a local feature extractor combining lip semantic segmentation network and convolutional neural network. The local feature extractor contains two branches inside. The first branch is composed of semantic segmentation network and convolutional feature extraction network, and the second branch is composed of convolutional feature extraction network. Then we concatenate the feature vectors obtained from the two branches to get the local lip feature vector.

2.1.1. Semantic segmentation network

Semantic segmentation is an important research field in image processing and computer vision. To achieve better segmentation accuracy, we use U-Net [3], which is a powerful network and suitable for lip region of interest segmentation. This is a very simple network, the first half of the network is used for feature extraction, the second half is up sampling. We use this network to segment the lip region and input it into the CNN feature extraction network to get the feature vector from the lip segmentation map.
2.1.2. Deep convolutional feature extractor

Lip feature extraction plays an important role in lip reading. Most researchers use deep convolutive neural network, such as ResNet [4], ShuffleNet [5], etc., while some others use SE-ResNet [6], which is a variant of ResNet, and achieve higher accuracy. To extract the feature more effective, we also use SE-ResNet as CNN feature extraction network to extract features from the input.

2.2. Adjacent feature extractor

To obtain the adjacent features of lips between adjacent frames more efficiently, we designed a graph structures to describe the change of the shape of lips based on the feature points marked by dlib toolkit [7]. The adjacent feature point graph consists of the feature points of three adjacent frames and some edges connecting the points in adjacent frames. The feature point graph is first transmitted to adjacent matrix and then feed it into a convolutional feature extraction network. Finally, the adjacent feature extractor output the adjacent feature vector.
2.3. Multi-feature fusion
The proposed local feature extractor and adjacent feature extractor will obtain two kinds of features vector, namely local and adjacent feature vectors. We will adopt three different level fusion strategies, namely feature-level, score-level, and decision-level fusion. For feature-level, we adopt the method of splicing feature vectors, for score-level, we adopt the method of summation, and for decision-level, we adopt the method of maximum voting.

3. Training data
3.1. Lip reading dataset
To demonstrate the superiority of our model, we design a more difficult dataset for lip reading. It contains 10 phrases that are the same as ouluvs2 [8]. There are 10 subjects, 5 male and 5 female, who all have very distinct facial features, such as thick beards and different skin tones. During the recording, we used a HD camera to record the video of the front face and added the light with obvious changes to simulate the real environment.

3.2. Lip region segmentation dataset
We will use part of the video in the lip reading dataset and split it into pictures and use LabelMe annotation tool to make lip region segmentation dataset for training the lip segmentation network.
4. Experiment and results

4.1. Pretrain the lip segmentation network
Our proposed local feature extractor needs to pretrain a lip segmentation network. We adopted dice loss as the loss function of U-Net segmentation network and optimized it with Adam optimizer. The learning rate was 1e-4, and train for 30 epochs.

4.2. Lip reading experiment
The proposed network had two independent feature extractors. First, we trained each feature extractor individually, using Adam optimizer with initial learning rate 1e-5. We use cosine learning rate schedular to adjust the learning rate. The back-end classifiers used for feature extractor training were all BiGRU. We train each extractor for 30 epochs. After that, the two feature extractors were combined through feature-level fusion, and then input into back-end classifier for classification. Then, 30 epochs were continued, and the learning rate was set to 1e-6. The segmentation network parameters are fixed throughout the training process. We compared the proposed feature extractor with traditional SE-ResNet+BiGRU model. The experiment result shows that the two-feature fusion achieves better accuracy than single-feature and traditional convolutional feature extractor.

| Front-end feature extractor       | Accuracy |
|----------------------------------|----------|
| SE-ResNet                        | 67.1     |
| Local feature extractor          | 72.3     |
| Adjacent feature extractor       | 74.5     |
| **Local feature extractor + Adjacent feature extractor** | **76.2** |

4.3. Multi feature fusion experiment
We compared various feature fusion methods, including feature-level, score-level, and decision-level. Three networks were trained separately, each train for 30 epochs using the Adam optimizer with initial learning rate 5e-6 and cosine learning rate schedular. During the whole training process, the parameters of the two feature extractors are fixed, and the model only updates the parameters of feature fusion and back-end BiGRU classifier. The experiment result shows that the feature-level fusion is the best choice of our network.

| Methods          | Accuracy |
|------------------|----------|
| Decision-level   | 73.1     |
| Score-level      | 74.7     |
| **Feature-level**| **76.2** |

5. Conclusion
In this paper, we propose two feature extractors, which can extract local features and adjacent features to effectively represent the lip feature when in wild environment. We also produced a very challenging dataset containing multiple factors affecting the accuracy of recognition. Experiments on this dataset show that our model is better to cope with the real environment than other models. We also studied a variety of feature fusion methods, among which feature-level is the most suitable for the proposed network. We hope that the proposed lip feature extractor could provide some valuable references to the related researchers.
References

[1] Xiao J, Yang S, Zhang Y, et al. Deformation Flow Based Two-Stream Network for Lip Reading. 2020.

[2] Petridis S, Wang Y, Ma P, et al. End-to-End Visual Speech Recognition for Small-Scale Datasets[J]. Pattern Recognition Letters, 2020, 131.

[3] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, 2015.

[4] Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.

[5] Ma N, Zhang X, Zheng H T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[J]. Springer, Cham, 2018.

[6] Squeeze-and-Excitation Networks. IEEE transactions on pattern analysis and machine intelligence, 2019.

[7] Davis E King, “Dlib-ml: A machine learning toolkit,” The Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[8] Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audio visual database for non-rigid mouth motion analysis," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, pp. 1-5