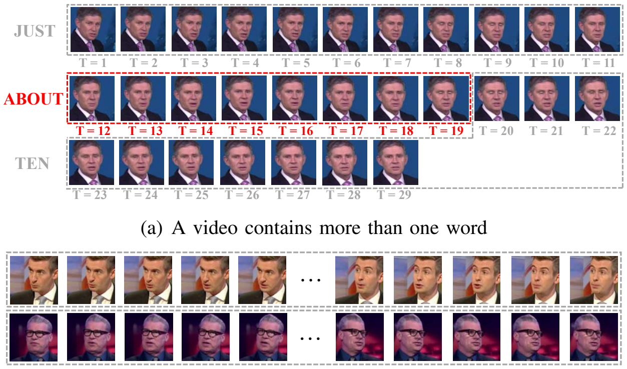
Mutual Information Maximization for Effective Lip Reading

Xing Zhao[1]*, Shuang Yang[2], Shiguang Shan[2,3], Xilin Chen[2,3]
[1] Zhejiang University of Technology, Hangzhou 310014, China
[2] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China

***Abstract*— Lip reading has received an increasing research interest in recent years due to the rapid development of deep learning and its widespread potential applications. One key point to obtain good performance for the lip reading task depends heavily on how effective the representation can be used to capture the lip movement information and meanwhile to resist the noises resulted by the change of pose, lighting conditions, speaker's appearance, speaking speed and so on. Towards this target, we propose to introduce the mutual information constraints on both the local feature's level and the global sequence's level to enhance the relations of them with the speech content. On the one hand, we require the features generated at each time step to carry a strong relation with the speech content by imposing the local mutual information maximization constraint (LMIM), so as to improve the model's ability to discover fine-grained lip movements and the fine-grained differences between words with similar pronunciation, such as "spend" and "spending". On the other hand, we introduce the mutual information maximization constraint on the global sequence's level (GMIM), to make the model be able to pay more attention to discriminate key frames related with the speech content, and less to various noises appeared in the speaking process. By combining these two advantages together, the proposed method is expected to be both discriminative and robust for effective lip reading. To verify this method, we evaluate on two large-scale benchmarks whose videos are collected from several TV shows with a wide coverage of the speaking conditions. We perform a detailed analysis and comparison on several aspects, including the comparison with the baseline of the LMIM and GMIM, and the visualization of the learned representation. The results not only prove the effectiveness of the proposed method but also report new state-of-the-art performance on both the two benchmarks.**

I. Introduction

Lip reading is a task to infer the speech content in a video by using only the visual information, especially the lip movements. It has many crucial applications in practice, such as assisting audio-based speech recognition [4], biometric authentication [2], aiding hearing-impaired people [22], and so on. With the huge success of deep learning based models for several related tasks in the computer vision domain, some works began to introduce the powerful deep models for effective lip reading in these three years [2], [18], [17], [14]. For example, [18] proposed an end-to-end deep learning architecture for word level visual speech recognition, which is a combination of convolutional networks and bidirectional Long Short-Term Memory networks, yielding an improvement of 6.8% on the accuracy than before. Besides the great impetus of deep learning technologies, several large-scale lip reading datasets, were released in recent years, such as LRW [6], LRW-1000 [23], LRS2 [5], LRS3 [1], and so on. These datasets have also contributed significantly to the recent progress in the development of lip reading.



(a) A video contains more than one word

(b) Two samples of "ABOUT"

Fig. 1. The word-level lip reading is a challenge task. **(a)** The actual frames of "ABOUT" include only frames at the time step T = 12∼19. **(b)** The same word label always have a greatly diversified appearances changes.

In this paper, we focus on the word-level lip reading, which is a basic but important branch in the lip reading domain. For this task, each input video is annotated with a single word label even when there are other pronunciations in the same video. For example, the sample in Fig. 1(a) is annotated as "ABOUT", but the actual frames of "ABOUT" include only frames at time step T = 12∼19, shown in the red boxes. The frames before and after this interval are corresponding to the word "JUST" and "TEN" respectively, not "ABOUT". This is consistent with the actual case where the exact boundary of a single word is always hard to get. This property requires a good lip reading model to be able to learn the latent regular patterns hidden in different videos of the same word label, and so to be able to pay more attention to valid key frames, but less to other unrelated frames. Besides the above problem of word boundary, the video samples corresponding to the same word label always have greatly diversified appearance changes, as shown in Fig. 1(b). All these properties require the lip reading model to be able to resist the noises in the sequence to capture the latent

* This work is done by Xing Zhao during his internship at Institute of Computing Technology, Chinese Academy of Sciences.

patterns which are consistent in various speech conditions.

In the meanwhile, due to the limited effective area of lip movements, different words probably give a similar appearance in the speaking process. Especially, the existence of homophones where different words may look the same or quite similar increases many extra difficulties to this task. These properties require the model being able to discover the fine-grained differences related to different word labels in the frame-level to distinguish each word from the other.

To solve the above issues, we try to introduce the mutual information maximization (MIM) on different levels to help the model learn both robust and discriminative representations for effective lip reading. On the one hand, the representation at the global sequence level would be required to have a maximized mutual information with the speech content, to force the model learning the latent consistent global patterns of the same word label in different samples, while being robust to the variations of pose, light and other label-unrelated conditions. On the other hand, the features at the local frame level would be required to maximize their mutual information with the speech content to enhance the word-related fine-grained movements at each time step to further enhance the differences between different words. By combining these two types of constraints together, the model could automatically find and distinguish the valid key frames corresponding to the target word, and ignore other unrelated frames. Finally, we evaluate the proposed approach on two large-scale benchmarks LRW and LRW-1000, whose samples are all collected from various TV shows with a wide variation of the speaking conditions. The results show a new state-of-the-art performance on both the two challenging datasets when compared with other related work in the same condition of using no extra data or extra pre-trained models.

The proposed method could also be easily modified to other existing models for other tasks, which may bring some meaningful insights to the community for other tasks.

II. Related Work

In this section, we provide an overview of the related literature on two closely related aspects, lip reading and mutual information based methods.

A. *Lip Reading*

When deep learning technologies are not so popular, many methods have achieved several encouraging results by using specifically-designed and hand-engineered features, such as optical flow [16], lip movement tracking, and so on. The classification is often done by Support Vector Machine [16] together with the Hidden Markov Models (HMMs) [3]. We refer to [24], [15] for a detailed review on these non-deep methods based lip reading. These previous work have provided an important impetus to the advancement of lip reading at the early stage.

With the rapid development of deep learning in recent years, more and more researchers gradually tend to perform the lip reading task by deep neural networks.

2D-CNN is the first type of network applied for lip reading to extract features for each frame. [12] proposed a system comprises a CNN and a hidden Markov model with Gaussian mixture observation model (GMM-HMM). The outputs of the CNN are regarded as visual feature sequences, and the GMM-HMM is applied for word classfication. In the later works [19], [5], long short-term memory (LSTM) or gated recurrent unit (GRU) is used to model the patterns on the temporal-dimensional issues. The CNN-LSTM based models, which can be trained in an end-to-end manner, has gradually become a processing pipeline for lip reading.

However, the mouth region at each frame is not consistent and the context in near frame plays an important role for effective lip reading. Several methods introduce 3D convolution to tackle this problem [14], [17], [23]. For example, LipNet [2] employed a 3D-CNN in the front-end on the visual frames for the lip reading task and obtained remarkable performance. Stafylakis et al. [18] combined 3D-CNN and 2D-CNN based networks to obtain features, which got a much higher accuracy on LRW dataset than before.

Besides directly applying different types of deep networks to lip reading, some recent impressive works begun to design particular models to solve the shortcomings of some existing networks for more effective lip reading. For example, Stafylakis et al. [17] utilized additional word boundary information to improve the performance on the word-level LRW dataset. [5] introduced the attention mechanism for selecting key frames in the sequence-to-sequence model. Wand et al. [20] improved the accuracy of lip reading by domain-adversarial training, which is expected to get speaker-independent features, which is beneficial to the final word classification. However, it is unable to apply in large scale dataset which contains over one thousand people. Recently, Wang [21] extracted both frame-level fine-grained features and short-term medium-grained features by a 2D-CNN network and 3D-CNN network respectively. In this paper, we propose a new way for effective lip reading. Specifically, we introduce the constraints on both the local feature level and the global representation level to make the model both be able to learn fine-grained features and pay attention to key frames respectively. At the same time, it is easy to train and doesn't increase many parameters while preserving good performance.

B. *Mutual Information Mechanism*

Mutual information (MI) is a fundamental quantity for measuring the relationship between two random variables. It is always used to evaluate the “amount of information” obtained about one random variable when given the other random variable. Based on this property, the mutual information of two random variables is always used as a measure of the mutual dependence between the two variables. Moreover, unlike the Pearson correlation coefficient which only captures the degree of linear relationship, mutual information also captures nonlinear statistical dependencies [9], and therefore has a wide range of applications.

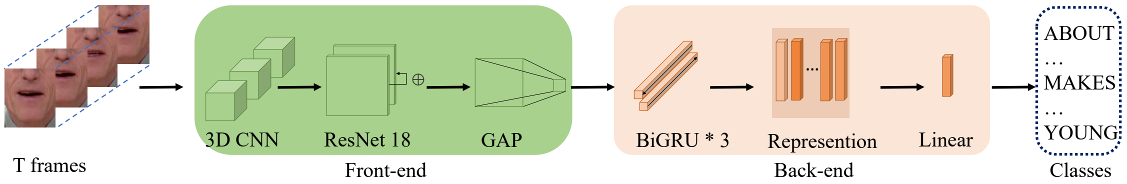


Fig. 2. Base architecture.

For example, Ranjay et al. [10] built a visual question generation model by maximizing the MI between the image, the expected answer and the generated question, leading to the model's ability to select corresponding features. Li et al. [11] tried to maximize the MI between the source and target sentences in the neural machine translation task to improve the diversity of translation results.

One work which has a bit relation with our work is Zhu et al. [25], who performed talking face generation by maximizing the MI between the words distribution and the facial/audio distribution. But in our work, we try to maximize the MI between the words distribution and the representation at different levels, to guide the model towards learning both robust and discriminative features for the lip reading task, which is totally different with [25].

III. THE PROPOSED MUTUAL INFORMATION MAXIMIZATION FOR LIP READING

In this section, we would first give an overview of the overall architecture. Then the particular manner to introduce mutual information mechanisms on different levels is presented. Finally, the optimization process to learn the model is provided.

A. The Overall Architecture

Let $\mathbf{X} = (x_1, x_2, ..., x_T)$ denotes the input sequence with T frames in total, where x_i is the feature vector of the i-th frame. The task of the model is to classify the input sequence into one of the C classes, where C is the total number of all the classes. Let $\mathbf{Y} = (0, 0, 1, ..., 0)$ denotes the annotated word label of the sequence, where $\mathbf{Y}$ is a $C-$dimensional one-hot vector with only a single 1 at the position corresponding to its word label index. We construct our base architecture with two principal components, front-end and back-end, which enable the total network to be trained end-to-end.

Specifically, given the input image sequence $\mathbf{X}$, a 3D-CNN layer is firstly applied on the raw frames, in order to perform an initial spatial temporal alignment in the sequence for effective recognition. A spatial max-pooling layer is then followed to compact the features in the spatial domain. It should be noted that we keep the temporal dimension unchanged in this procedure to avoid a further shortage of the movement information in the sequence because the duration of each word is always very short. In the next step, we divide the features into T parts and employ ResNet18 at each time step $t = 1, 2, ..., T$ to separately extract discriminative features. To improve the ability to capture fine-grained movements related to the spoken word, we impose the mutual information constraint on these features with the annotated label to enhance their relations in the learning process. Then all these features would be fed into a global average pooling(GAP) layer to compress the final features into $T \times D$ where D is the channel of the last layer and 512 in this paper, GAP is always used to reduce the dimensions of a tensor. We named the above module as the **Front-end**, including a 3D-CNN layer, a spatial pooling layer, a ResNet18 network, and a GAP layer, as shown in Fig. 2.

With the initial representation from the Front-end, a 3-layer Bi-GRU is followed to capture the latent patterns of the sequence in the temporal dimension. Bi-GRU contains two independent single directional GRUs. The input sequence is fed into one GRU as the normal order, and into another GRU as the reverse order. The outputs of the two GRUs would be concatenated together at each time step to represent the whole sequence. The output of the Bi-GRU is expected to be a global representation of the whole input sequence with dimension $T \times 2N$, where N is the neurons in each GRU. The representation will be finally sent to a linear layer for classification, we named these parts as the **Back-end**, as shown in Fig. 2. To improve its ability to resist noises and selecting key frames in the sequence, we impose another mutual information constraint on this global representation.

B. Local Mutual Information Maximization (LMIM)

As stated in the previous section, the performance of lip reading is heavily affected by the model's ability to capture the local fine-grained lip movements, so as to generate discriminative features to distinguish the words from each other, especially the homophones. MI-based constraint is a promising tool for learning good features in an unsupervised way, because we never need any additional data to train it. In this paper, we would introduce LMIM (Local Mutual Information Maximization) on ResNet18 to help the model focus more on the spatial related regions at each time step and produce more discriminative features. Especially when it comes to lip reading, the local features related to mouth regions are significant, so that the Front-end should perceive the variations of the local regions properly. Therefore, unlike most existing work [10], [25], we maximize the MI on each patch of the features rather than the whole features.

The process of the LMIM is shown in Fig. 3. We assume the feature map in the last layer of ResNet18 (which will be sent to the GAP layer) as $\mathbf{F}$ with has a shape of $H \times W \times D$,

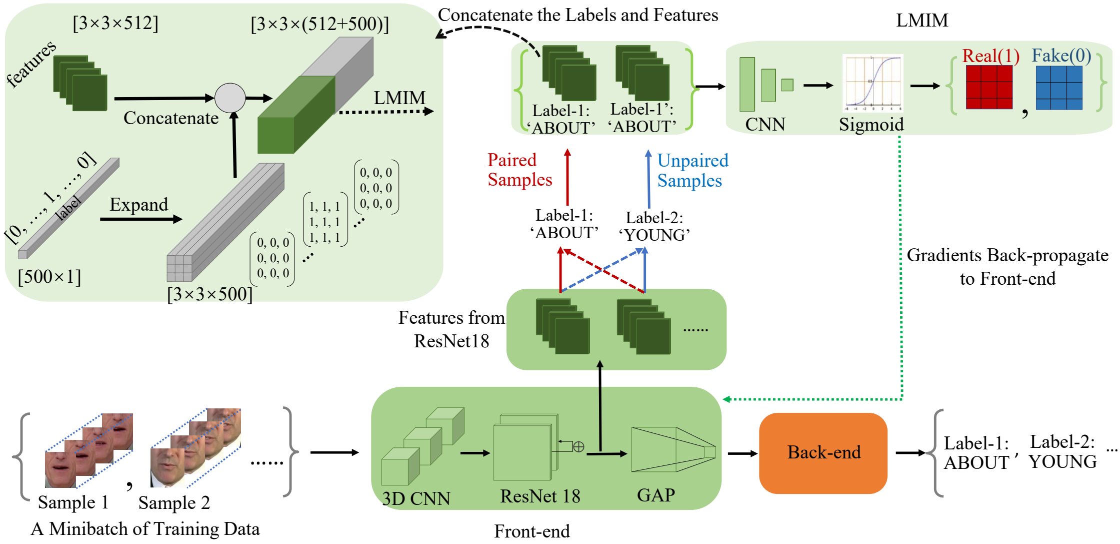


Fig. 3. The process of training the base network with the proposed LMIM. The total loss is computed by averaging all of the time steps and patches. The gradients from the LMIM will be backpropagated to the Front-end through the features sampled from the ResNet18. The LMIM will be dropped after training.

where H,W and D are the height, width and the channels respectively.

Because mutual information is notoriously hard to compute for unknown distribution, we estimate it with the help of deep network here. Following the representation of Jensen-Shannon(JS) MI estimator [8], [13]:

$$\begin{aligned}\widehat{\mathscr{I}}_{\theta}^{(JSD)}(A,B) = &\mathbf{E}_{p(A,B)}\left[-\varphi\left(-T_{\theta}(a,b)\right)\right] \\ &-\mathbf{E}_{p(A)p(B)}\left[\varphi\left(T_{\theta}(a,b)\right)\right],\end{aligned} \tag{1}$$

where $\varphi(k) = log(1+e^k)$, A and B are the two variables that we want to estimate the MI between them, T_θ is a continuous function that we directly use a network to approximate it. The $p(A,B)$ is the joint distribution of paired samples $\{a,b\}$, and the $p(A)p(B)$ is the marginal distribution of the unpaired samples $\{a,b\}$ by randomly sampling A and B. In the optimization process, because $\varphi(k) = log(1+e^k)$ is a monotone increasing function, so maximizing the JS MI estimator is equivalent to optimize (1) with $\varphi(k) = log(1+k)$ when the formula is equal to the binary cross-entropy loss.

We assume the feature $\mathbf{F}$ have $H \times W$ local patches $(f_1, f_2, ..., f_{H\times W})$ which looks like we separate the original image to $H \times W$ patches when the receptive field of the features are mapped to the original image. The label of each sample is expanded from the one-hot vector of dimension $C \times 1$ to the same height and width as $C \times H \times W$ by repetition. Then we concatenate the labels and features together to obtain a representation of dimension $(C+D) \times H \times W$, which would be used as the input to compute the Local Mutual Information Maximization network (LMIM). To obtain the local mutual information in each position of the $H \times W$ locations, we employ two convolutional layers with kernel size 1×1 on the concatenated representation. Then a sigmoid activation is applied to the last layer to simulate the value of the mutual information. Please note that the architecture of the network in this step can be any other form, because it is just applied to approximate a continuous function T_θ. The output layer must be based on a sigmoid activation function when using the binary cross-entropy. The dimension of the outputs of the LMIM is $H \times W$, each number explains how much the corresponding patch is related with the given word label. When the inputs are paired samples, we expect the mutual information of every patch close to 1 (Real). In other cases, we expect it to 0 (Fake). For sampling unpaired samples, we randomly concatenated the features with the labels in the same batch.

Therefore, the optimization for LMIM can be denoted as a binary cross-entropy loss:

$$\begin{aligned}L_{(LMIM)} = &E_{p(F,Y)}\left[\log\left(LMIM\left(f,y\right)\right)\right] \\ &+E_{p(F)p(Y)}\left[\log\left(1-LMIM\left(f,y\right)\right)\right].\end{aligned} \tag{2}$$

Noting that in this stage, we don't solve any temporal issues, the features of T time steps in an input video will be sent to LMIM successively. In the end, we compute the mean of the loss at all time steps for computing the gradients.

C. Global Mutual Information Maximization (GMIM)

In each sequence, the amount of valuable information provided by different frames is not equal. In LRW dataset and several practical cases, there are many frames in a sequence that is not relevant to the given target word label. However, the popular way in current related work is to average all the time steps to get the final representation, which has neglected this point and has superior performance when come to practice.

To solve this problem, we introduce global mutual information maximization on the global representation obtained from the Bi-GRU. Specifically, we introduce an additional

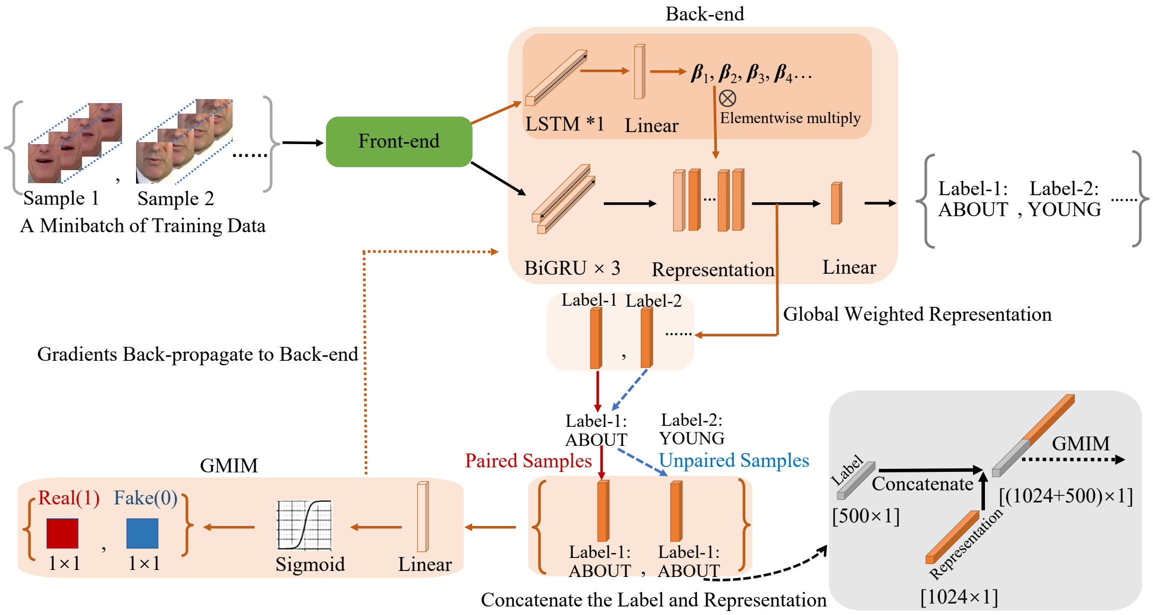


Fig. 4. The process of training the network with the proposed GMIM, noted that when we apply the GMIM, a single layer LSTM and a linear layer are also added to the Back-end for computing the weight of each frame, it will be retained after training while the GMIM will be dropped.

LSTM together with a linear layer on the outputs of the Front-end. This additional LSTM would allocate different weights β for different frames according to the target word. The total architecture is shown in Fig. 4.

Based on the outputs $\mathbf{Z}(T \times 2N)$ of the 3-layer Bi-GRU layers and the weighted value $\beta(T \times 1)$, the final global representation is the weighted average of the outputs $\mathbf{Z}$ at all time steps:

$$\mathbf{O} = \frac{\sum_{t=1}^{T} \beta_t \cdot \mathbf{Z}_t}{T}. \tag{3}$$

The output $\mathbf{O}(2N)$ will be sent to a linear layer to transform its shape from $2N$ to C, C is the total number of classes, the final representation of the whole sequence $\mathbf{O}(C)$ is applied to get the classification score as

$$\hat{Y}_i = \frac{\exp(O_i)}{\sum_{j=1}^{C} \exp(O_j)}. \tag{4}$$

However, the weights β would not be normalized in some manner. For related valuable key-frames, the weight should be positive and can be of any value. While for unrelated frames, we just want its weight close to zero, not a negative number for the optimization problem. Therefore we use ReLU to obtain the weight β as

$$\beta_t = \mathrm{ReLU}\left(\mathbf{W}_{linear} \times \mathbf{LSTM}(\mathbf{G})_t + \mathbf{b}_{linear}\right), \tag{5}$$

where $\mathbf{G}$ is the outputs of the GAP layer, $\mathbf{W}_{linear}$ and $\mathbf{b}_{linear}$ are the parameters of the linear layer and $\mathbf{LSTM}(\mathbf{G})_t$ denotes the hidden state at time step t of the extra LSTM layer.

To guide the learning of the weights, we constrain the weighted average vector to contain most of the information about the target word. Specifically, we maximize the MI between the weighted average representation $\mathbf{O}(2N)$ and label $\mathbf{Y}$, both of which will be fed into the global mutual information maximization network (GMIM), which consists of two linear layers and outputs a scalar after a sigmoid activation. If the inputs come from paired samples, we expect the outputs of GMIM as large as possible and even close to 1 (Real). In other cases, the output is expected to close to 0 (Fake). So the objective function can be written as:

$$\begin{aligned} L_{(GMIM)} = & E_{p(O,Y)}\left[\log\left(GMIM(o,y)\right)\right] \\ & + E_{p(O)p(Y)}\left[\log\left(1 - GMIM(o,y)\right)\right]. \end{aligned} \tag{6}$$

D. Loss Function

Combining the cross-entropy loss with the LMIM and GMIM optimization function, the final objective loss function for the whole model is:

$$L_{total} = -\sum_{i=1}^{C} Y_i \log \hat{Y}_i - L_{(LMIM)} - L_{(GMIM)}, \tag{7}$$

where the first term is the cross-entropy loss and Y_i is the label. These three cross-entropy loss have the similar number, there is no need to allocate different weights to each loss, which makes the networks can be trained easily and stably.

IV. Experiments

In this section, we first evaluate the performance of our modified architecture (baseline) which can be trained easier than others. Then we conduct an ablation study on the proposed LMIM and GMIM (GLMIM) and figure out how they help the model get better results respectively. we also evaluate the performance of the baseline with the GLMM and compare it with other state-of-the-art lip reading methods on the two largest benchmark datasets. Finally, we visualize the discriminative representations leaned with the GLMIM. Code and models will be available at `https://github.com/xing96/MIM-lipreading`.

A. Datasets

We evaluate our method on two large-scale word-level lip reading benchmarks, LRW and LRW-1000. The samples in both of these two datasets are collected from TV shows, with a wide coverage of the speaking conditions including the lighting conditions, resolution, pose, gender, make-up etc.

LRW [6]: is released in 2016, contains 500 classes with more than a thousand speakers. It is the largest English database for word-level audiovisual speech recognition, which has 500 classes and displays substantial diversity in speech conditions. The training set instances reach to 488766, validation and test set each contains 25000 instances. LRW remains a challenging dataset and therefore has been widely used by most existing deep learning based methods, even the accuracy is hard to rise by one percentage point.

LRW-1000 [23]: The dataset is the largest and also the only one for Mandarin lip reading dataset, which has 1000 classes. The training set contains 603097 instances and test set contains 51578 instances, totaling 57 hours. Samples of the same word are not limited to a previously specified length range to allow the existence of various speech rates, which brings the more difficulties. This dataset aims at covering a "natural" variability over different speech modes and imaging conditions to incorporate challenges encountered in practical applications.

B. Implementation Details

The inputs for network are all cropped or resized to 88 × 88. The kernel size, stride and padding of the first 3D-CNN are $(5,7,7)$, $(1,2,2)$ and $(2,3,3)$ respectively. Each GRU or LSTM layer has 1024 neurons (which means each Bi-GRU contains 2048 neurons). The Adam optimizer is used with weights decay 0.00005. In the training process, the learning rate would decay from 0.0001 to 0.00001 when the accuracy doesn't increase. Dropout is utilized at Bi-GRU layers to mitigate the overfitting problem.

C. Baseline

We adopt [14] as the base architecture. The accuracy of our re-implementation on LRW is a little lower than the value in the original paper. We use the modified version as described in III-A and take it as our baseline when using no MI constraint. As is shown in Table I, our modified architecture is superior to the base architecture, which achieves 82.14% accuracy on LRW dataset.

TABLE I

COMPARISON OF THE MODIFIED BASELINE.

Method	Accuracy
Petridis[14]	82.00%
Petridis[14](our re-implement)	81.70%
The Modified Baseline Architecture	**82.14%**

TABLE II

IMPROVEMENT ON WORDS WITH SIMILAR PRONUNCIATION

Class	Baseline	Baseline with LMIM	Improvement
MAKES	62%	74%	**12%**
MAKING	80%	92%	**12%**
POLITICAL	82%	90%	**8%**
POLITICS	84%	92%	**8%**
STAND	48%	60%	**12%**
STAGE	70%	80%	**10%**
NORTH	78%	90%	**12%**
NOTHING	78%	86%	**8%**
SPEND	36%	46%	**10%**
SPENDING	78%	82%	**4%**

D. Effect of the LMIM

In order to illustrate the effectiveness of the proposed LMIM, we train the baseline network and the baseline network with the LMIM separately. The LMIM will be dropped after training, which means that these two networks have the same architecture and parameters. When we compare the accuracy between these two networks, we find that the network trained with the LMIM performs better, the accuracy have risen by more than a percentage point. We conduct the statistics of the accuracy of each class, after applying the LMIM, most of the classes with high accuracy improvement are similar words, such as MAKES/MAKING and POLITICAL/POLITICIANS, as is shown in Table II. The LMIM helps the baseline architecture to extract the local fine-grained features, which is significant to improve the ability to distinguish the words which have similar pronunciation.

E. Effect of the GMIM

Selecting key frames is essential because a video contains more than one word, this is why we apply GMIM to make the model pay different attention to all frames. We directly choose the model trained with LMIM in IV-D as the contrastive network because of its excellent ability to extract fine-grained features. For the sake of fairness, the Front-end is fixed and only the Back-end is trained with GMIM. Without sending any additional word boundary information,

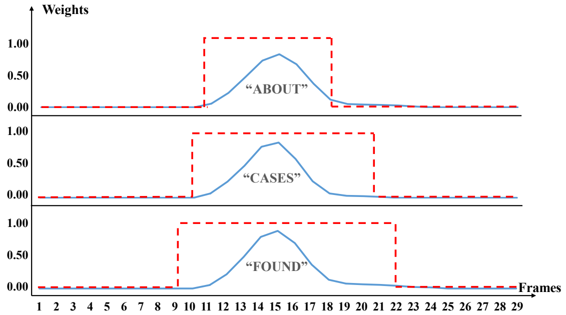


Fig. 5. We randomly sample three words and show the weights of each frame learned with GMIM. The blue line shows the learned weight for each frame. The red dashed line denotes the word boundary for the target word when its value is 1.

TABLE III

Comaprison with other related work on LRW.

Method	Accuracy
Chung[2018][7]	71.50%
Chung[2017][5]	76.20%
Petridis[2018][14]	82.00%
Stafylakis[2017][18]	83.00%
Wang[2019][21]	83.34%
Baseline	82.14%
Baseline+LMIM	83.33%
The Proposed GLMIM	**84.41%**

TABLE IV

Comaprison with other related work on LRW1000.

Method	Accuracy
LSTM-5	25.76%
D3D[2018][23]	34.76%
3D+2D	38.19%
Wang[2019][21]	36.91%
Baseline	**38.35%**
Baseline+LMIM	**38.69%**
The Proposed GLMIM	**38.79%**

the model learns the key frames precisely and the accuracy has increased by one percentage point. After we fine-tune the total model with the proposed GLMIM, we get a new state-of-the-art result.

The weights learned with the proposed GMIM, as is shown in Fig. 5. The horizontal axis represents the 29 frames of the video and the vertical axis represents the numeric of the learned weights. The blue line shows the learned weights for each frame. The red dashed line denotes the word boundary for the target word when its value is 1. Meanwhile, the context information is important for lip reading, if we only use the frames which are among the word boundary, the performance would drop off, which has been experimented in [17]. Our model trained with GMIM not only learns the key frames successfully and pays more attention to the frames which are included in the word boundary, but also allocates small amount of weights to the frames close to the word boundary for capturing the context information.

F. Compare with state-of-the-art methods

We apply the proposed GLMIM to our main architecture and compare it with the current state-of-the-art methods on LRW. Although our baseline is not the best, but after we apply the LMIM, the accuracy rises about 1.21%, our proposed LMIM can help the CNN to capture more discriminative and accurate features for the main task. Meanwhile, the GMIM can help our main architecture to select key frames rather than averaging all time steps directly. When we apply the LMIM and GMIM together, the accuracy of our model reaches 84.41%. Comparing with other lip reading methods which have no additional inputs except the visual information, as shown in Table III, we get the best result and provide a new state-of-the-art result on the LRW dataset.

LRW1000 is a new dataset, which has a large variation of speech conditions including lighting conditions, resolution, speaker's age, pose, gender, and make-up, etc. The best result is only 38.19% up to now. It is challenging to obtain a good performance on this dataset while we achieve a high accuracy of 38.79% which outperforms the state-of-the-art results. Table IV gives the accuracy of our models. The improvement of the GMIM is smaller when comparing with the improvement on LRW, this interesting phenomenon may be due to the useless frames in LRW-1000 is less than LRW, even in some cases, there is no additional frames for other words which makes the context information unavailable.

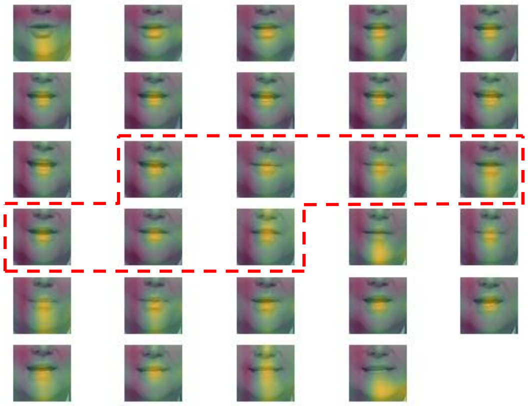
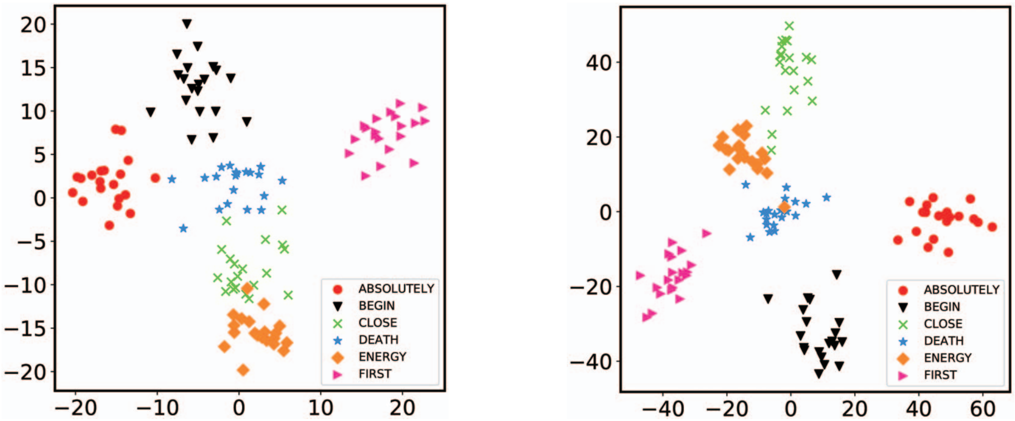


Fig. 6. Visualization of the features of the word "ABOUT", the network concentrates on the regions around the mouth.

G. Visualization

For better demonstrating the ability of our method, we choose a sample word "ABOUT" and visualize the feature maps as shown in Fig. 6. The word boundary of the target word "ABOUT" are surrounded by the red dashed line. We performed summation over the features in the last layer of ResNet18 along the channel dimension and mapped the values (after normalizing) to different colors. Then we draw the color to the same regions in the original input frame according to the receptive field. The bright yellow regions corresponding to the saliency regions with big feature values. We can see that, the learned features automatically highlight the mouth related regions.

We want to figure out how the proposed GLMIM achieve high results, we choose 6 classes and each of them contains 20 samples. We send them separately to our original baseline architecture and the architecture with applying the proposed GLMIM. We extract the final representations **O** which will be sent to the linear layer for classification. We apply PCA to reduce its dimension form higher dimensions to 2 dimensions for better visualization. As is shown in Fig. 7, the variance among these classes before applying GLMIM ranges only from -20 to 20; While the variance has been enlarged to the interval between -40 and 60 after applying GLMIM, which means the variance among the classes have been greatly increased due to the introduction of the proposed GLMIM, which makes it easier to distinguish different classes.



(a) Results before applying the GLMIM

(b) Results after applying the GLMIM

Fig. 7. PCA visualization of the final represention form the Bi-GRU. With the help of the GLMIM, the architecture gets more discriminative results.

V. Conclusion

In this paper, we propose the mutual information maximization based methods both for local fine-grained feature extraction and global key frames selection. We also modify the existing model for lip reading that make it can be trained easier, and we get the best results on the two largest word-level lip reading datasets. Lip reading is still a challenge task, but it's worth to do because of its great value.

VI. Acknowledgments

This work is partially supported by National Key R&D Program of China (No. 2017YFA0700804) and National Natural Science Foundation of China (No. 61702486, 61876171).

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[2] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[3] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar. The natural statistics of audiovisual speech. *PLoS computational biology*, 5(7):e1000436, 2009.

[4] T. Chen and R. R. Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, 2017.

[6] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[7] J. S. Chung and A. Zisserman. Learning to lip read words by watching videos. *Computer Vision and Image Understanding*, 173:76–85, 2018.

[8] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[9] J. B. Kinney and G. S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014.

[10] R. Krishna, M. Bernstein, and L. Fei-Fei. Information maximizing visual question generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2008–2018, 2019.

[11] J. Li and D. Jurafsky. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*, 2016.

[12] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Lipreading using convolutional neural network. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[13] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[14] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552. IEEE, 2018.

[15] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, 22:23, 2004.

[16] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. C. Azemin, and J. Gubbi. Lip reading using optical flow and support vector machines. In *2010 3rd International Congress on Image and Signal Processing*, volume 1, pages 327–330. IEEE, 2010.

[17] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos. Pushing the boundaries of audiovisual word recognition using residual networks and lstms. *Computer Vision and Image Understanding*, 176:22–32, 2018.

[18] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[19] M. Wand, J. Koutník, and J. Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, 2016.

[20] M. Wand and J. Schmidhuber. Improving speaker-independent lipreading with domain-adversarial training. *arXiv preprint arXiv:1708.01565*, 2017.

[21] C. Wang. Multi-grained spatio-temporal modeling for lip-reading. *arXiv preprint arXiv:1908.11618*, 2019.

[22] K. Xu, D. Li, N. Cassimatis, and X. Wang. Lcanet: End-to-end lipreading with cascaded attention-ctc. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 548–555. IEEE, 2018.

[23] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. *arXiv preprint arXiv:1810.06990*, 2018.

[24] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen. A review of recent advances in visual speech decoding. *Image and vision computing*, 32(9):590–605, 2014.

[25] H. Zhu, A. Zheng, H. Huang, and R. He. High-resolution talking face generation via mutual information approximation. *arXiv preprint arXiv:1812.06589*, 2018.