KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://karyailham.com.my/index.php/arca
2462-1927

Enhanced Model Compression for Lip-reading Recognition based on Knowledge Distillation Algorithm

Hu Qian[1,*], Kuryati bt Kipli[1], Tengku Mohd Afen[1], Liu Yuan[2], Liu Xiangju[2], Wang Bo[2]

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), Kota Samarahan, Sarawak, Malaysia
[2] School of Computer and Information Engineering Qilu Institute of Technology Jinan, Shandong, China

ARTICLE INFO

Article history:
Received 27 November 2026
Received in revised form 20 February 2026
Accepted 15 April 2026
Available online 4 May 2026

Keywords:
lip-reading, model compression, deep learning, knowledge distillation

ABSTRACT

Lip-reading is the process of understanding what a speaker is saying by observing changes in the speaker's mouth. With the development of computer vision and natural language processing technology, lip-reading recognition technology has been paid more and more attention and applied. Especially in recent years, deep learn-based lip-reading recognition technology has made great progress, and the lip-reading model is becoming increasingly complex, requiring a lot of computing resources, and it is difficult to apply to portable mobile devices directly. The lip-reading recognition model LipPC-Net proposed in this paper is built with a large Chinese lip-reading data set based on Chinese phonetic rules and grammatical features and consists of two main parts: the P2P sub-model and the P2C sub-model. The P2P sub-model is a model for identifying pinyin sequences from pictures, while the P2C sub-model is a model for identifying Chinese character sequences from pinyin. However, due to China's rich and ambiguous language features, its training and optimization rely on a graphics processor (GPU), which has high computing power and storage space requirements, so staying in the theoretical research stage and large-scale promotion and application is difficult. To realize the transformation of scientific research results as soon as possible, highlight the universality of the lip-reading model under the intelligent environment, and solve the problem of how to embed the lip-reading model into the mobile terminal with limited computing and storage capabilities, this paper proposes three knowledge distillation compression algorithms to complete the compression of the Chinese character sequence output by the model. They are an offline model compression algorithm based on multi-feature transfer (MTOF), an online model compression algorithm based on adversarial learning (ALON), and an online model compression algorithm based on consistent regularization (CRON). MTOF can solve the problem of single migration features, ALOF middle layer feature mutual learning is ignored, CRON can solve the problem of decision boundary fuzzy features are ignored, through the three compression algorithms to fit and learn the transformation between different features, so that portable mobile terminals with limited hardware resources can carry this model. Then, realize the practical application value of assisting the communication of the deaf and mute.

* *Corresponding author.*
E-mail address: 21010379@siswa.unimas.my

https://doi.org/10.37934/arca.43.1.97110

1. Introduction

Lip-reading recognition, which can be done by observing the speaker's mouth changes, "reading" or "partially reading" what it says, is the process of understanding lip movements.

Currently, the mainstream lip-reading recognition technology is based on the lip-reading recognition of computer vision and natural language processing technology. This lip-reading recognition method uses computer vision technology to identify speech content directly from the video image of someone speaking. Firstly, the face is continuously identified from the image, the person speaking is judged, and the characteristics of this person's continuous mouth changes are extracted. Then, the continuously changing features were input into the lip-reading recognition model to identify the corresponding pronunciation of the speech population type. Then, based on the recognized sounds, the most likely natural language statement was calculated. This kind of lip-reading recognition has the advantages of being non-contact, low cost, and the user does not need to contact the equipment, health, and safety, so it is more acceptable. Therefore, more and more researchers have been working on the lip-reading recognition system of this kind of method.

In recent years, with the continuous maturity of artificial intelligence and neural network technology, deep learning-based lip-reading recognition technology has also been proposed. Compared with traditional lip-reading recognition, deep learning-based lip-reading recognition integrates lip detection and feature extraction into one process and greatly improves the accuracy of lip-reading recognition and the performance of the lip-reading recognition model.

Although the research on lip-reading recognition based on deep learning has achieved good results, the practical application of deep learning in lip-reading recognition still faces many challenges. Firstly, to cope with complex tasks, the network model of deep learning is continuously deepened and complicated. For example, from the five layers of the early LeNet model, the current general purpose.

The ResNet series model has been developed to up to 152 layers. Meanwhile, the number of model parameters has also increased dramatically, from tens of thousands of parameters in the early days to several million. The training and deployment of these models require a lot of computing resources and are difficult to apply directly to the currently popular embedded and mobile devices. In addition, the Chinese lip-reading recognition model is very large due to the complexity of Chinese characters and the existence of a large number of homophones and near-phonics in Chinese. Therefore, its training and optimization process generally relies on the GPU server, which takes at least 4-5 days to train. The model will occupy more than 3G of video memory if used even longer. However, if the lip-reading recognition model is built using large Chinese data sets such as CAS-VSR-Wlk (formerly LRW-1000) [1] or Chinese Mandarin Lip-reading (CMLR) [2], it will require longer training time and occupy larger video memory. Therefore, it is still difficult to embed the lip language model into mobile devices with limited computing and storage capabilities to achieve universal application and give full play to its real practical value. Considering this practical problem, this paper will explore the relevant model compression methods.

In response to the shortcomings of deep learning in the current industry, Hinton et al. first proposed Knowledge Distillation (KD) [3] in 2015, which uses a complex deep network model to transfer knowledge to a shallow small network model. The advantage of this learning model is that it can reuse the existing model resources and use the information contained in it to guide the new training stage. The cross-domain application also changes the dilemma that the previous task or scene changes need to re-make the data set and training model, greatly saving the cost of deep neural network training and application.

This paper proposes a method to compress the Chinese lip-reading model through a knowledge distillation algorithm so that general terminal equipment can run and load the lip-reading recognition model to realize the transformation of scientific research results from the theoretical stage to practical application and then realize the practical application value of lip-reading recognition technology.

In summary, our key contributions are as follows:

- A Chinese lip-reading recognition model, LipPC-Net, was constructed according to Chinese pronunciation rules and grammatical features.
- With the Knowledge Distillation (KD) algorithm, the model compression problem is solved by learning and fitting between features, and then the compression application of LipPC-Net is realized.
- The future development direction of lip-reading recognition and model compression is discussed.
- Summarize this paper

2. Related Work

2.1 Lip-reading

In recent years, with the continuous maturity of artificial intelligence and neural network technology, deep learning-based lip-reading recognition technology has also been proposed. Compared with traditional lip-reading recognition, deep learning-based lip-reading recognition integrates lip detection and feature extraction into one process and greatly improves the accuracy of lip-reading recognition and the performance of the lip-reading recognition model. In 2014, Noda et al. achieved 56% accuracy in words and 44.5% accuracy in phrases by pre-training the hybrid model of VGGNet [4] and Recurrent Neural Network (RNN) [5]. In 2016, Chuang et al. proposed a spatiotemporal convolutional neural network on the BBCTV dataset, which achieved good word classification accuracy [6]. In the same year, Wand et al. introduced the LSTM recurrent neural network on the GRID [7] dataset and achieved 79.6% lip-reading recognition accuracy [8]. In 2017, Assael et al. proposed the LipNet model, which achieved 79.6% accuracy on GRID datasets [9]. In 2019, Margam et al. used mixed 2D and 3D convolutional neural networks for feature extraction and then used bidirectional LSTM for classification, obtaining 98.70%-word accuracy on the GRID dataset [10]. In 2019, Shuang Yang et al. proposed a large-scale benchmark CAS-VSR-Wlk for naturally distributed field lip-reading. In the three years following Shuang Yang et al., many researchers have used deep learning methods to study this data set continuously, and the accuracy of lip recognition has increased from 36.91% to 53.82%.

2.2 Knowledge Distillation

The core idea of the model compression algorithm based on knowledge distillation is to use the teacher-student approach, train the teacher model with superior performance in advance, then fix its parameters and put forward its features to help improve the performance of the student model.

The concept of knowledge distillation was proposed by Bulica in 2006 and summarized and developed by Hinton in 2014, the core idea of which is first to train a complex network model. This complex network's output and the accurate data labels are then used to train a smaller network. Hence, knowledge distillation frameworks typically contain a complex model (called the Teacher model) and a smaller model (called the student model). By knowledge distillation, model accuracy

can be improved, model delay can be reduced, network parameters can be compressed, and domain migration between labels can be solved.

The knowledge distillation algorithm is divided into two different ways: offline knowledge distillation and online knowledge distillation. Table 1 compares offline knowledge distillation with online knowledge distillation. These two compression methods are described below.

2.2.1 Off-line Knowledge Distillation

The offline knowledge distillation algorithm uses the "teacher-student" mode. It guides the pre-trained teacher model to help improve the performance of the student model and achieve the purpose of model compression.

Bucilua et al. [11] first used the Soften label in 2006 to soften knowledge distillation and feature migration, compensating for the problems of Hard label with weak signal supervision and low information entropy. Later, Hinton et al.

Table 1
Comparison Of Off-Line Knowledge Distillation and Knowledge On-Linedistillation

Year	Reference	Compression method	New ideas proposed	Feature transfer
2015	Adriana et al. [12]	Offline Knowledge Distillation	Deeper and narrower student models	unidirectional
2017	Zagoru Yko et al. [14]	Offline Knowledge Distillation	The attention mechanism of the convolutional network	unidirectional
2018	Kim et al. [15]	Offline Knowledge Distillation	convolution operation	unidirectional
2019	Huang et al. [16]	Offline Knowledge Distillation	A distribution matching Problem	unidirectional
2019	Heo et al. [17]	Offline Knowledge Distillation	The transfer knowledge of the teacher model	unidirectional
2019	Shen et al. [18]	Offline Knowledge Distillation	Adversarial learning Strategies	unidirectional
2017	Hou et al. [22]	Online Knowledge Distillation	DualNet	bi-directional
2018	Zhang et al. [19]	Online Knowledge Distillation	Deep Mutual Learning (DML)	bi-directional
2018	Lan et al. [27]	Online Knowledge Distillation	A multi-branch network framework based on the peer-to-peer model	bi-directional
2019	Gao et al. [20]	Online Knowledge Distillation	The online collaborative learning model	bi-directional
2020	Chung et al. [23]	Online Knowledge Distillation	Introduced adversarial learning mechanism	bi-directional
2020	Kimet et al. [29]	Online Knowledge Distillation	Used the Ensemble Logits of the model	bi-directional

Guidance layers but narrower width to learn to fit the middle layer representation features of their teacher model and the final output results. Thus, a student model with better generalization ability and faster running speed can be obtained. Yim et al. [13] proposed a new idea of "teaching them to fish," taking the inner product between the features of two different network layers as the transfer feature to guide the student model to learn from the teacher model. Zagoruyko et al. [14] proposed the attention mechanism of the convolutional network, using the characteristic

information of attention to guide students to imitate the teacher model with good performance and improve the performance of the student model. Kim et al. [15] used convolution operation to assist the student model in "understanding" and "translating" the transferred knowledge information in the teacher model. Huang et al. [16] regarded knowledge distillation as a problem of distribution matching. By designing the maximum mean difference loss function based on feature distribution, they realized the distribution matching of neuronal selection modes between teacher and student models. They achieved the goal of knowledge distillation and transfer. Heo et al. [17] proposed to use the transfer knowledge from the teacher model to assist in the initialization of the student model to obtain better initialization parameters for the student model. In addition, some researchers proposed adding "antagonism" into the distillation model, that is, introducing discriminators to guide students to fit the output of the teacher model, which also achieved good results. For example, Shen et al. [18] used adversarial learning strategies to train and optimize the student model and promoted discriminators to distinguish the characteristics of teachers and students. Thus, it can guide students to learn the model and fit the characteristic distribution process of teachers' models.

Although these algorithms have performed well, in the traditional off-line knowledge distillation algorithm, selecting and pre-training the teacher model consumes extra computing resources and training time. In addition, when the student model is trained, the parameters of the teacher model are fixed and can only be used to output the feature distribution of the specified layer. Therefore, the feature transfer process is unidirectional. In other words, the teacher model cannot learn timely feedback from the student model, and its performance is restricted to a certain extent.

2.2.2 Online Knowledge Distillation

Considering the limitations of the offline distillation algorithm, Zhang et al. [19] proposed a deep mutual learning (DML) model in 2018 to simplify the model's training process further. The core idea is to train and guide peer models to learn from each other and fit each other's output prediction patterns simultaneously to achieve the common improvement of model performance.

Researchers have successfully proposed various improved models based on the DML algorithm framework. For example, Gao et al. [20] and Anil et al. [21] proposed the online collaborative learning model, mainly used to guide the same untrained network model to learn the target task simultaneously. Each model learns the average classification probability from the peer model; Hou et al. [22] proposed DualNet, which fused two separate classifiers into one fusion classifier. In the training process, the separate classifier and the fusion classifier learned from each other to conduct local optimization and global optimization, respectively. Chung et al.[23] introduced adversarial learning mechanism to guide peer models to learn from each other the loss value of knowledge distillation based on feature graph. Other algorithms [24-25] are mainly based on the output of the specified model to periodically update other network weight parameters to achieve online video distillation. Song et al. [26] and Lan et al. [27]built a multi-branch network framework based on the peer-to-peer model [28], in which all structures and the shared backbone layer can build a single network model together, and any target node network in the whole multi-branch can be optimized. Although collaborative learning and multi-architecture can promote online knowledge distillation, not all models can obtain helpful information. Therefore, Kim et al. & Lin et al. [29-30] propose to use the Ensemble logs of the model to establish a dynamic teacher or group leader and then distill it back into all peer-to-peer networks. The learning of the model is enhanced in a closed loop.

The online distillation algorithm is based on mutual learning between peer models compared to the traditional two-stage offline knowledge distillation training process. It can complete the training

in one stage. It saves computing resources and training time. It can realize the effective compression application of the model with superior performance, so the suitable model can be selected for deployment and application on the mobile device.The teacher-student model does not constrain this algorithm research and has a wider application field.

3. Methodology

This paper proposes a model compression for Chinese lip-reading recognition based on a knowledge distillation algorithm. The flow chart of the whole working process is shown in Fig.1. Firstly, on the premise of the Chinese data set, to overcome the ambiguity existing in the direct translation from picture sequence to sentence, the Chinese lip recognition task was divided into two sub-tasks before building the lip language model, and the respective neural network models were used for separate pre-training. That is, the pinyin sequence recognition sub-model P2P from lip picture to pinyin and the Chinese character sequence recognition sub-model P2C from pinyin to Chinese character. After the two sub models have been pre-trained and converged, they are combined to carry out the end-to-end recognition model LipPC-Net for the picture-to-Chinese character sequence. In the near future, experiments can prove that the stepwise training optimization strategy will be very effective in dealing with Chinese lip recognition problems and improving the accuracy of model recognition. Then, the Chinese character sequence output by the LipPC-Net model is successively compressed by the MTOF algorithm, ALON algorithm, and CRON algorithm to achieve the purpose of embedding and optimization on mobile devices.

3.1 Data Preprocessing

Before constructing the lip-reading model, the data should be preprocessed correctly because the data preprocessing results will directly affect the performance results. In the preprocessing step, the redundant information is discarded to focus on the region of interest (ROI). The quality of ROI extraction also affects recognition performance. For this purpose, the Viola-Jones detector first detects all images from the sequence [31].

The faces in each sequence were then aligned using the detected landmarks and normalized to the population mean and variance. We then apply affine transformations to extract a mouth-centered clipping from the frame. The resulting frames were adjusted to a fixed size of 224 × 224, and then all the same sequence of frames were randomly flipped horizontally to expand the data set further. Fig.2 shows the process of image preprocessing. To increase the robustness of the model, we changed the motion speed by deleting and copying frames, with a probability of 0.05 per frame. These steps ultimately need to be evaluated in advance through several settings.

3.2 Construction of Chinese lip-reading recognition model

This study must build a Chinese lip-reading data set according to Chinese pronunciation rules and grammatical features. The pronunciation rules of Chinese mainly include two aspects: pinyin and syllable. Chinese Pinyin comprises 23 consonants, 24 vowels, and four tones. Initial consonants, or consonants, are used before vowels to form a complete syllable together with vowels and tones. Syllables are the basic components of language and are the smallest units that can be distinguished from speech information by hearing. The number of Chinese characters is far greater than the number of syllables.

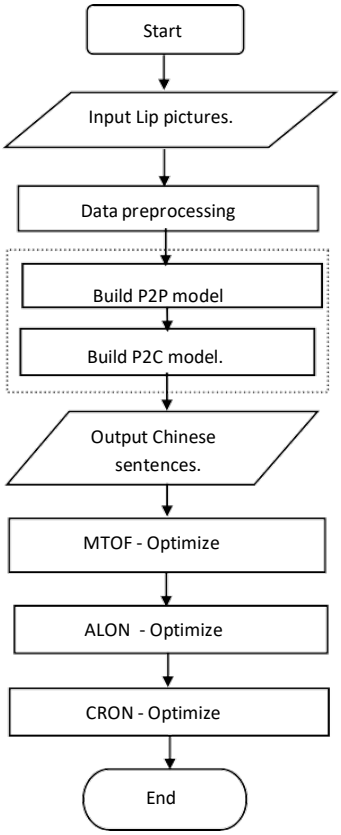

Fig. 1. Research methodology flowchart.

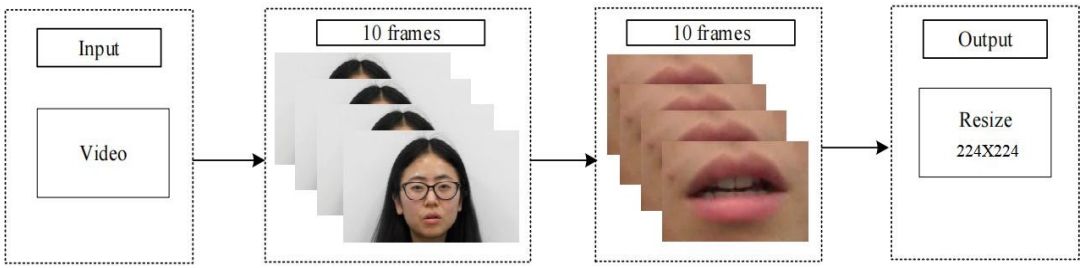

Fig. 2. The process of image preprocessing

Chinese has many homophones and polyphonic, and different characters correspond to the same mouth shape. Therefore, Chinese is a language with great ambiguity. Therefore, to overcome the fuzziness of direct translation of pictures into Chinese characters, the Chinese lip-reading recognition task in the dataset was first split into picture-to-pink-word (P2P) and pink-to-Chinese character (P2C), and different network models and corresponding training techniques and strategies were used to train and optimize these two subtasks. When P2P and P2C sub-models converge, they are combined into the Chinese lip-reading recognition model LipPC-Net to start the end-to-end training mode. This strategy can automatically match long sequences of lip pictures and natural sentence segmentation and improve lip language models' robustness.

3.2.1 Picture to Pinyin Recognition Submodel (P2P)

Fig.3 shows the identification submodel P2P from picture to pinyin sequence. Different from the ordinary image feature extraction task, lip-reading recognition needs to capture the subtle changes of lips in the speaking process, and judge the speaker's speaking content through the continuity and difference of the lip pictures extracted by the feature.

The P2P sub-model is mainly a deep neural network complex model constructed by VGG-M[32] and LSTM. VGG-M has good classification performance on the ImageNet[33] data set and occupies relatively small memory space, which makes training more convenient. Therefore, in the process of image recognition and pinyin sequence recognition, VGG-M with a full connection layer removed was first selected as the shallow network for lip feature extraction of the Chinese data set. It is a hybrid structure including 3D- CNN and 34-layer ResNet. ResNet18 is the most lightweight and is often used for lightweight feature extraction tasks at the word level [34]. As one of the most commonly used sentence-level lip-reading feature extraction networks, the ResNet34 model can show the best comprehensive performance. Since this study focuses on the lip-reading recognition of Chinese sentences, the ResNet34 model can ensure the integrity and richness of features. Since the changes of lips are sequential, the pre-processed lip images cannot effectively capture the time information after removing the VGG-M of the full connection layer, so two layers of LSTM are selected to extract the time information between the feature maps generated by 34 layers of ResNet, to ensure the integrity and richness of the features.

In optimizing the P2P sub-model, the Connectionist Temporal Classification (CTC) loss function is used to complete automatic alignment and error calculation. The CTC loss function is mainly used to solve the problem that inputs and outputs need to be aligned. Compared with traditional algorithms, such as the hidden Markov model, this function has better performance.

3.2.2 Pinyin to Chinese Character Recognition sub-model (P2C)

Fig.4 shows the pinyin to Chinese character sequence recognition sub-model P2C. It consists of a gated recursive unit GRU sensor embedded with an attention mechanism. As pointed out by Chan et al.[35], when the time step is too large, it is difficult to thoroughly train the sequence-to-sequence recognition model. Therefore, the P2C sub-model can be divided into two language modules, encoder, and decoder, to overcome the ambiguity in pinyin and Chinese character sequence mapping. In the encoder and decoder modules, two independent attention mechanisms, Attentione and Attention, are used to enhance the model training, where Attention is only given the feature vector v and the state vector s, that is, during the entire LipPC-Net training process. The encoder and decoder are pre-trained with auxiliary training data before performing end-to-end sequence-to-sequence joint optimization.

In the training process of Encoder, input data and labels are pinyin sequences. The encoder first sets pinyin sequences C=c1, c2, … ci, which is transformed into the embedded space, and then the output vector sequence O=o1, o2, … oi is calculated by weighted connection. Each output vector oi used in the parametric next may enter the predicted distribution of Pr(ci+1|oi). The modeling goal of Encoder is to maximize the total logarithmic probability value of the training sequences,

$$\sum^{i-1} \log \Pr\left(c_{i+1} \mid c \le i\right) \tag{1}$$

In the training process of the Decoder, input data and labels are Chinese character sequences, and the attention vector and output state are mixed to generate the Chinese character vector rk, which contains the required information to generate the output of the next time step. Finally, the

output probability distribution of Chinese characters is calculated through the full connection layer and Softmax, such as public notice (1). In each time step, k, tk, gk, and rk are the output, state, and context vectors, respectively, and ek is the output value. Finally, parameters retained by Encoder and Decoder during separate training are used to initialize the P2C sub-model.

The attention mechanism is introduced in the sub-model P2C because it can significantly improve the recognition accuracy of the sequence model. If the attention mechanism is not introduced, the model will forget previously helpful information and produce sentence outputs independent of the input sequences.

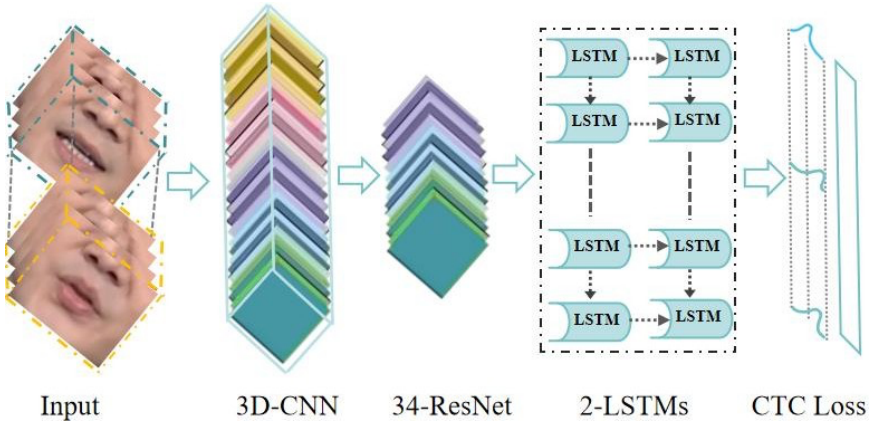

Fig. 3. Picture to Pinyin Recognition Submodel P2P

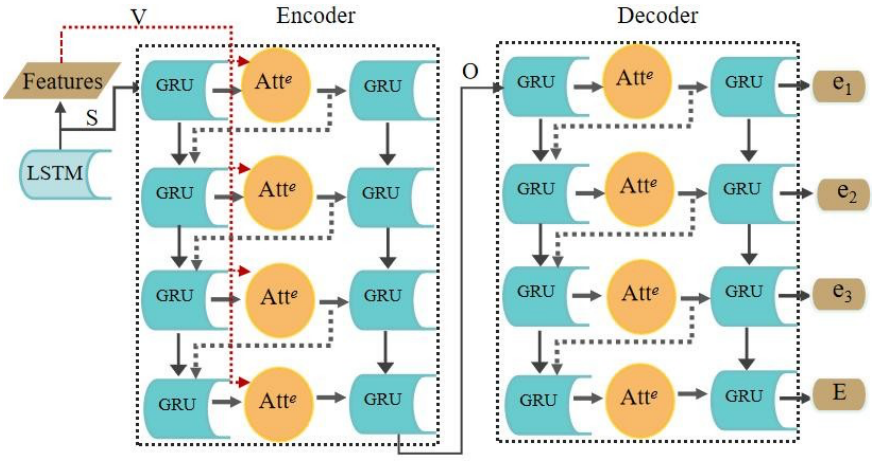

Fig. 4. Picture to Pinyin Recognition Sub-model P2C

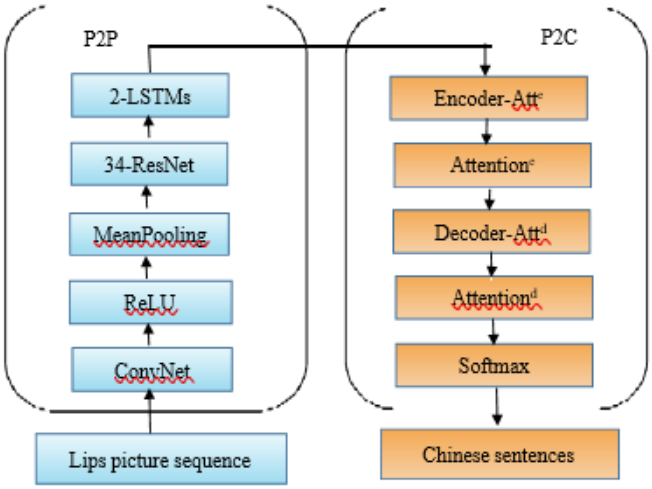

Fig. 5. Lip recognition model LipPC-Net

$$t_k, g_k = GRU(g_{k-1}, e_{k-1}, r_{k-1}) \tag{2}$$

$$r_k = t.Attention(t, g_k) \tag{3}$$

$$P(e_i \mid e_{<i}) = soft\max f(t_k, r_k)) \tag{4}$$

3.2.3 Lip reading recognition model LipPC-Net

After the Pinyin sequence recognition sub-model P2P and the Chinese character sequence recognition sub-model P2C are pre-trained respectively, they are combined to form the end-to-end lip recognition model LipPC-Net, as shown in Fig.5. After the lip sequence images are input to LipPC-Net, they are jointly processed end-to-end by the P2P and P2C sub-models, and the corresponding Chinese sentences are directly output. In the joint optimization training process of LipPC-Net, the CTC loss function of the P2P sub-model is removed, and its output is the input of the P2C submodel; the output feature vectors of each LSTM cell are input one after another into the encoder GRU cell in the P2C submodel, and at the same time, the state vector s and feature vector v are also input into the GRU and Attention unit.

3.3 Model Compression

As the Chinese lip model, LipPC-Net is limited by the computing resources and storage capacity of mobile intelligent devices; it will not be able to be used in practice, etc., the following three knowledge distillation compression algorithms are proposed in this section.

3.3.1 Off-line model compression algorithm based on multi-feature transfer (MTOF)

MTOF algorithm can solve the problem of a single migration feature. Fig.6 is the flow chart of the MTOF framework. The orange arrow represents the pre-trainingprocess of the teacher model in the first stage, the green arrow represents the training optimization of the studentmodel in the second stage with the assistance of the transfer feature, and the gray arrow is the training technique added inthe model training process, that is, the output of the convolutional layer and the output of the deconvolution layerof the corresponding dimension are directly added and then input to the next module. First, the algorithm uses a complementary objective task to pre-train the teacher model to be feature-rich. Then, the complementary adversarial function

and feature loss function are used to guide thestudent model to fit the multi-feature information of the teacher model from both spatial and pixel dimensions of the image, respectively, to improve the utilization of migrated features, and then improve the feature extraction ability of the student model to make it easy to deploy while alsoapproximating the performance of the teacher model to achieve the purpose of model compression. Under the condition of generating effective accuracy, MTOF can implement LipPC-Net compression applications to a certain extent.

3.3.2 Online model compression algorithm based on adversarial learning (ALON)

ALON algorithm can solve the problem of neglected mutual learning of intermediate layer features. Fig.7 is the flow chart of the ALON framework. The algorithm adopts a result-driven algorithm to achieve mutual fitting between the final prediction distributions of the peer-to-peer model on the one hand, and introduces a complementary process-driven algorithm on the other hand, i.e., it uses discriminators and alignment containers to guide the chunked modules in the model to learn the intermediate layer outputs at the same location in the peer-to-peer network as well as the feature distributions at the highest layer in an adversarial learning and pixel fitting manner, respectively, to improve the utilization rate of the features inside the model and the interaction rate between internal and higher-level features, thus speeding up the convergence of the model and enhancing the robustness and performance of the peer-to-peer model. With the same compression rate, ALON can further optimize the recognition accuracy of LipPC-Net and its simplified model on the CAS-VSR-Wlk data set.

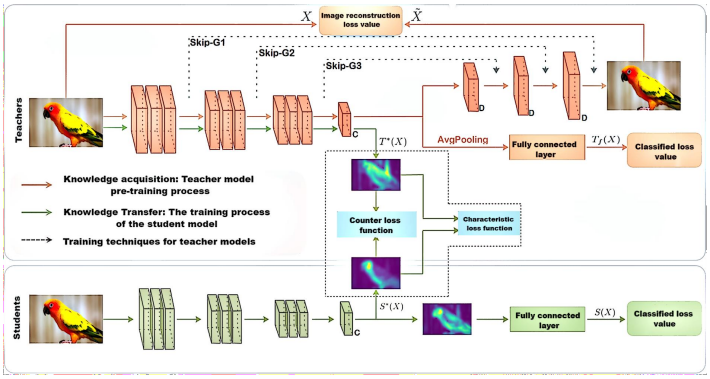

Fig. 6. MTOF Framework Flow Chart

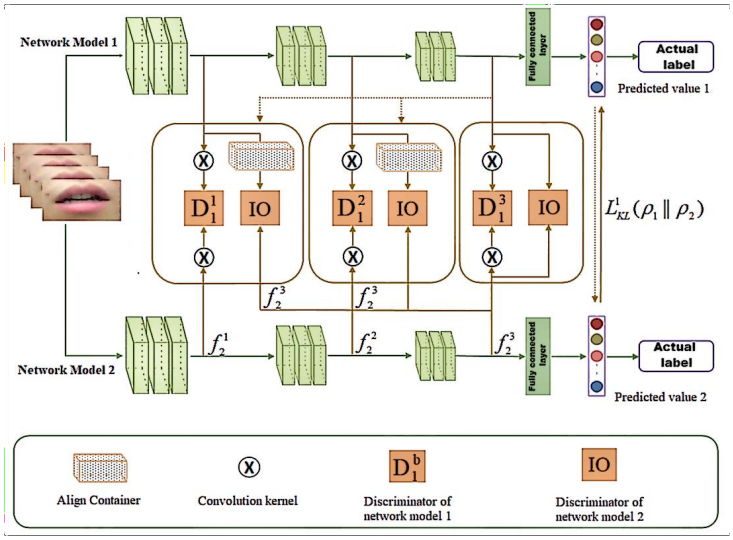

Fig. 7. ALON Framework Flow Chart

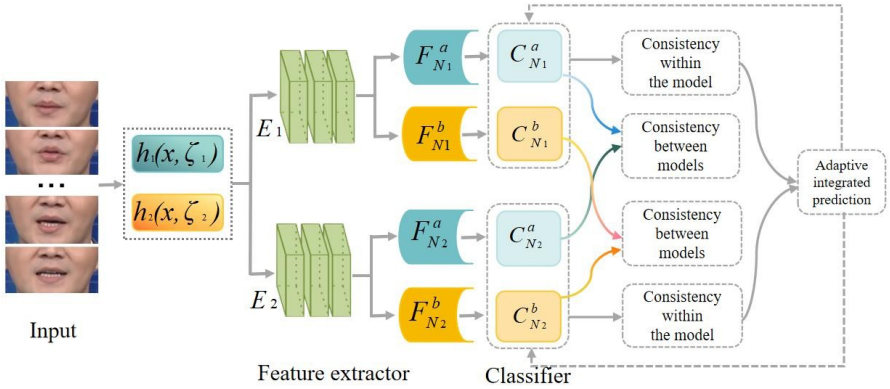

Fig. 8. CRON Framework Flow Chart

3.3.3 Online model compression algorithm based on consistent regularization (CRON)

CRON algorithm can decide the problem of boundary fuzzy features being ignored. Fig.8 is the flow chart of the CRON framework. Each model in CRON consists of a shared feature extractor and a pair of task-specific classifiers. classifiers of the same model and the corresponding classifiers of different models between different classifiers of the same model and between the corresponding classifiers of different models to measure intra-model consistency and inter-model consistency, respectively. These two types of consistency are jointly used in the training and updating the feature extractor to improve the extraction ability of fuzzy features. In addition, the intra-model consistency is used with the average output of each model to generate the final integrated prediction, which guides all classifiers to learn from each other to improve the discrimination of fuzzy features. The partially simplified model trained by CRON can generate recognition accuracy similar to that of LipPC- Net, laying a theoretical foundation for embedding LipPC- Net into mobile portable applications.

MTOF, ALON, and CRON compression algorithms can realize the compression application of LipPC-Net on the Chinese lip-reading data set to a certain extent, optimize the simplified model in turn, and improve the performance and recognition accuracy of the Chinese lip-reading recognition model. Therefore, the experimental and theoretical basis for applying the Chinese lip recognition model is laid.

4. Development and Challenge

In the era of artificial intelligence, lip-reading recognition and model compression technology has been developed rapidly. Only when the compressed lip-reading recognition model is embedded in the mobile device terminal and applied to the actual life scenes such as medical rehabilitation, smart home, and assisting criminal investigators in arresting criminals can its specific application value be truly realized. Through techniques such as knowledge distillation, model pruning, and lightweight model building, it is possible to compress the number of model parameters by hundreds or even thousands of times without losing model performance. In the coming years, model compression techniques will continue to evolve to be more efficient and accurate to accommodate growing data and computing demands. For example, some new technologies, such as adaptive compression and reversible compression, are becoming research hotspots.

The three model compression algorithms MTOF, ALON, and CRON discussed in this paper, while ensuring the accuracy of Chinese lip-reading recognition, can achieve the purpose of compression, simplification, and embedding of LipPC-Ne model into mobile terminals, but it may also affect some functions or characteristics of the model. For example, if you compress the model too much, you may compromise the model's generalization ability, making it difficult to adapt to different data distributions. Furthermore, if the model is over-compressed, it may slow down the operation speed, especially when processing large amounts of data, which may affect the performance of real-time lip recognition. If the model is over-compressed, it may also impair the ability to recognize dynamic features in lip-reading.

4. Conclusions

This paper proposes a knowledge distillation-based lip-reading recognition framework. The lip-reading recognition model LipPC-Net adopts a knowledge distillation compression algorithm for compression optimization, which can solve the problem of the computing resources and storage capacity of mobile intelligent devices being limited and not being able to be used on the ground. Future work will be done to implement the proposed model on existing Chinese data sets such as CAS-VSR-Wlk and CMLR to measure accuracy, precision, recall, robustness, and interpretability performance. In addition, verify whether the three knowledge distillation algorithms.

MTOF, ALON, and CRON can aim to embed the lip-reading model into mobile devices and optimize it to improve its robustness and stability.

Acknowledgment

This work was mainly done during the firstauthor's doctoral studies at the Sarawak University of Malaysia.

References

[1] Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., … Chen, X. (2019). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019. Institute of Electrical and Electronics Engineers Inc.

[2] Ya Zhao, Rui Xu, and Mingli Song. A Cascade Sequence-to- Sequence Model for Chinese Mandarin Lip Reading. ACM International Conference on Multimedia in Asia 2019.

[3] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[4] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. (pp. 1–14). International Conference on Learning Representations, ICLR.

[5] Noda K., Yamaguchi Y., & Nakadai K. (2014). Lip reading using convolutional neural network. Made Available by the Northern Territory Library Via the Publications Action, 25(6): 1840-1849.

[6] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (Vol. 2017-January, pp. 3444–3450). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CVPR.2017.367.

[7] Cooke M., Barker J., & Cunningham S., et al. (2006). An audiovisual corpus for speech perception and automatic speech recognition. Acoustical Society of America, 120(5): 2421-2424.

[8] Wand, M., Koutník, J., & Schmidhuber, J. (2016). Lipreading with long short-term memory. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (Vol. 2016-May, pp. 6115–6119). Institute of Electrical and Electronics Engineers Inc.

[9] Assael Y. M., Shillingford B., & Whiteson S., et al. (2016). Lipnet: Sentence-level lipreading[EB/OL]. https://arXiv.org/abs/1611.01599.

[10] Margam, D. K., Aralikatti, R., Sharma, T., Thanda, A., Roy, P. AK. S., & Venkatesan, S. M. (2019), ‘‘Lipreading with 3D-2D-CNN BLSTM-HMM and word-CTC models,’’ 2019, arXiv:1906.12170. [Online]. Available: http://arxiv.org/abs/1906.12170.

[11] Bucilua, C., Caruana, R., Mizi,l A. N. (2006) Model compression. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, United States, 2006: 535-541.

[12] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). FitNets: Hints for thin deep nets. In 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

[13] Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (Vol. 2017-January, pp. 7130– 7138). Institute of Electrical and Electronics Engineers Inc.

[14] Zagoruyko, S., & Komodakis, N. (2017). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

[15] Kim, J., Park, S. U., & Kwak, N. (2018). Paraphrasing complex network: Network compression via factor transfer. In Advances in Neural Information Processing Systems (Vol. 2018-December, pp. 2760–2769). Neural information processing systems foundation.

[16] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (Vol. 2017-January, pp. 2261 – 2269). Institute of Electrical and Electronics Engineers Inc.

[17] Heo, B., Lee, M., Yun, S., & Choi, J. Y. (2019). Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019 (pp. 3779–3787). AAAI Press.

[18] Shen, Z., He, Z., & Xue, X. (2019). MEAL: Multi-Model ensemble via adversarial learning. In 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019 (pp. 4886–4893). AAAI Press.

[19] Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep Mutual Learning. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 4320– 4328). IEEE Computer Society.

[20] Gao, L., Lan, X., Mi, H., Feng, D., Xu, K., & Peng, Y. (2019). Multistructure-based collaborative online distillation. Entropy, 21(4).

[21] Anil, R., Pereyra, G., Passos, A., Ormandi, R., Dahl, G. E., & Hinton, G. E. (2018). Large scale distributed neural network training through online distillation. In 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

[22] Hou, S., Liu, X., & Wang, Z. (2017). DualNet: Learn Complementary Features for Image Recognition. In Proceedings of the IEEE International Conference on Computer Vision(Vol. 2017-October, pp. 502–510). Institute of Electrical and Electronics Engineers Inc.

[23] Chung, I., Park, S. U., Kim, J., & Kwak, N. (2020). Feature-map-level online adversarial knowledge distillation. In 37th International Conference on Machine Learning, ICML 2020 (Vol. PartF168147-3, pp. 1984–1993).

[24] Cioppa, A., Deliege, A., Istasse, M., De Vleeschouwer, C., & Van Droogenbroeck, M. (2019). ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (Vol. 2019-June, pp. 2505–2514).

[25] Mullapudi, R. T., Chen, S., Zhang, K., Ramanan, D., & Fatahalian, K. (2019). Online model distillation for efficient video inference. In Proceedings of the IEEE International Conference on Computer Vision (Vol. 2019-October, pp. 3572–3581). Institute of Electrical and Electronics Engineers Inc.

[26] Song, G., & Chai, W. (2018). Collaborative learning for deep neural networks. In Advances in Neural Information Processing Systems (Vol. 2018-December, pp. 1832–1841).

[27] Lan, X., Zhu, X., & Gong, S. (2018). Knowledge distillation by on- the-fly native ensemble. In Advances in Neural Information Processing Systems (Vol. 2018-December, pp. 7517–7527).

[28] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 07-12-June-2015, pp. 1–9).

[29] Kim, J., Hyun, M., Chung, I., & Kwak, N. (2020). Feature fusion for online mutual knowledge distillation. In Proceedings - International Conference on Pattern Recognition (pp. 4619–4625). Institute of Electrical and Electronics Engineers Inc.

[30] Lin, R. J. X., Fan, J. (2019). Mod: A deep mixture model with online knowledge distillation for large scale video temporal concept localization[EB/OL].

[31] P. Viola, M. Jones. Rapid Object Detection using a Boosted Cascade of Simple. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.

[32] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In BMVC 2014.

[33] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei: ImageNet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. 248–255.

[34] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 2016- December, pp. 770–778). IEEE Computer Society.

[35] Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (Vol. 2016-May, pp. 4960–4964). Institute of Electrical and Electronics Engineers Inc.