Lip Reading Using Multi-Dilation Temporal Convolutional Network

Binyan Xu[1] and Haoyu Wu[2]

[1]*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China*
[2]*School of Physics, Xi'an Jiaotong University, Xi'an, China*
xby233@stu.xjtu.edu.cn

Abstract

In recent years, lip reading has attracted extensive research attention as deep learning shows great potential in computer vision. In this work, we proposed Multi-Dilation Temporal Convolutional Networks (MD-TCN) to predict individual words in lip reading tasks. Although Temporal Convolutional Networks (TCNs) have lately demonstrated promising importance in a variety of video sequence tasks, ordinary TCNs' bottom layers still have tiny receptive fields and are unable to reproduce complicated temporal dynamics in scenarios of lip-reading tasks. To tackle this problem, we use dual dilated convolution in the network instead of typical dilated convolution to capture more powerful temporal features. Furthermore, our method incorporates a self-attention block after each convolutional layer to further enhance the classification and screening capabilities of the model. On the lip-reading in the wild (LRW) dataset, our MD-TCN Model achieves 85.7 percent accuracy and is an effective method for individual word prediction.

Keywords

Lip-reading, Temporal Convolutional Networks, Visual Speech Recognition, Self-Attention

1. Introduction

Lip Reading, also known as Visual Speech Recognition, is a task of recognizing words based on the movement of the speaker's lips without audio assistance. Due to the complexity of Lip Reading, human lip readers need to take a long period of professional training to ensure the accuracy, which usually requires a high cost. Machine Lip Reading has become a very hot topic in video processing due to its relative low cost and even higher accuracy than human lip readers. And with the development of informatization, a real-time and fast lip-reading solution is required in many scenarios. Especially for speech recognition in a high-noise audio signal environment, the fusion of video signal and speech signal can greatly improve the accuracy of recognition, and the robustness of the system can also be greatly improved. At the same time, because of the nature of the lip-reading task, it is easy to apply the model to other video recognition tasks, such as action recognition and emotional semantic analysis.

This paper will build an end-to-end Visual Speech Recognition model and test it on both English and Mandarin data to better evaluate the model performance. Researchers generally divide Visual Speech Recognition into two steps: the front-end structure of visual feature extraction and the back-end structure of time series information recognition [1]. For visual feature extraction, VGG or Resnet are usually used as feature extraction tools in deep learning [2]. Because the lip-reading dataset usually has a large time series scale, 3D-CNN is also used for auxiliary data compression [3]. For time-series information recognition, deep learning generally adopts a series of time-series models, such as Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) networks, and Gated Recurrent Units (GRU) [4]. In recent years, the use of Temporal Convolutional Networks (TCN) in the back-end structure of Lip Reading has also become an outstanding scheme due to its good performance in many natural language processing tasks. Attention mechanisms have also become a popular model for video processing as they continue to prove their high effectiveness in the Natural Langrage Processing (NLP) and Computer Vision (CV) domains. However, these methods still have some limitations, such as poor robustness and low accuracy. This paper mainly adopts a network structure with 3D-CNN + dense-Resnet18 as the front end and Self-Attention Temporal Convolutional Networks SA-TCN as the back end, which got a good result on LRW dataset.

This article will be divided into five main sections. The second part of the article will review the classic literature on Visual Speech Recognition and the deep neural network architecture used in this paper, such as discussing previous research on Visual Speech Recognition and temporal convolutional networks. The third part of the article will focus on the feature extraction model of 3D-CNN + dense-Resnet18 and the sequence model of TCN, and discuss their advantages. The fourth part of the article provides an in-depth discussion of the experiments, introducing the dataset, experimental setup, experimental results, and a discussion of the results. Finally, the fifth section of the article will conclude by discussing the limitations of this study and directions for future research.

2. Related Works

2.1. Lip Reading

Before the deep learning gold rush, Lip Reading was mostly accomplished by depending on manually derived features, such as Discrete Cosine Transform (DCT) [2], Support Vector Machines (SVM) [3], and Hidden Markov Models (HMM) [1] etc. With the rapid advancement of deep learning, an increasing number of scholars have attempted to tackle the Lip Reading task using deep learning approaches in recent years. In 2014, Noda et al. [4] first proposed to apply the Convolutional Neural Network (CNN) to the Lip Reading task. This paper used 2D-CNN as the feature extractor to extract the lip feature vector, and it uses HMM as the backend to complete the classification. In subsequent work [5][6], Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) gradually replaced HMM as the mainstream back-end classifiers. LipNet [7] is the first approach to extract spatiotemporal features using 3D-CNN and present them to BGRU for classification. Stafylakis et al. [8] suggested a network topology that employed a 2D residual network on top of a 3D convolutional layer as the front-end (LSTM as the back-end), and they obtained a big accuracy breakthrough on the LRW dataset [11]. In recent years, Martinez and Ma et al. [9] proposed to apply Temporal Convolutional Networks (TCN) to the Lip Reading task, and it achieved good accuracy. In addition to the above-mentioned direct application of different network structures to Lip Reading, many researchers now begin to design some unique modules to achieve better results. In 2018, Stafylakis et al. [13] improved word-level lip-reading performance on the LRW dataset by extracting word boundary information. Chung et al. [14] applied an attention mechanism to select keyframes for sequence-to-sequence models. The word-level lip reading accuracy has reached 88.5% [9] and 55.7% [10] on the LRW [11] dataset and the LRW-1000 [12] Mandarin dataset, respectively.

2.2. Temporal convolutional networks (TCN)

Though RNN-like neural networks such as GRU and LSTM have been widely employed for time series applications, sequential models with higher parallelism and faster training have also received extensive attention in recent years. Lea et al. [15] first proposed Temporal Convolutional Networks (TCNs) for video action segmentation. The encoder and decoder of this network are both two-layer one-dimensional convolutional blocks. In 2018, Bai et al. [16] suggested an effective and simple TCN structure that surpassed RNN models in a variety of time series problems. Martinez et al. [17] first proposed Multi-scale TCN architecture to mix up long-term and short-term information, which improved robustness of network over time domain. Although the Temporal Convolutional Network introduced in [16] is a causal one-way model, it can also be adapted to an acausal structure in practical classification tasks (such as our lip-reading task). The work of [9] adopts a non-causal TCN structure and introduces densely connected layers to improve the performance of the network on complex datasets. However, these models may not consider the autocorrelation within the series. Many studies [18, 19] have shown that considering the autocorrelation of sequential models can effectively increase the prediction accuracy. As a result, we propose a TCN-embedded temporal self-attention strategy to increase the capture of sequence autocorrelations.

3. Methodology

3.1. Overview

The main framework of our technique is depicted in Fig. 1. The input is a raw video from a dataset with the shape $B \times T \times H \times W$, where T denotes the temporal dimension and H, W denotes the input video's height and width, respectively. After we adapt a face detection to input video, it can easily be transformed and cropped to gray-scale mouth Region of Interests (RoIs). We start by using a 3D convolutional layer to approximately extract the spatial-temporal features with the shape of $B \times T \times H_1 \times W_1 \times C_1$, where H_1 and W_1 are the modified height and width, and C_1 is the feature channel number, as described in [17]. We apply a 2D ResNet-18 [20] on top of this layer to generate features with the form $B \times T \times H_2 \times W_2 \times C_2$. To summarize the information of lip characteristics and compress the dimension to $B \times T \times C_2$, the following layer uses spatial average pooling. After the pooling layer, the temporal dynamics are modeled using our suggested Multi-Dilation TCN (MD-TCN). To complete the temporal information into C_4 channels, the output tensor ($T \times C_3$) is routed through another average pooling layer. The Softmax layer that follows predicts single word probability.

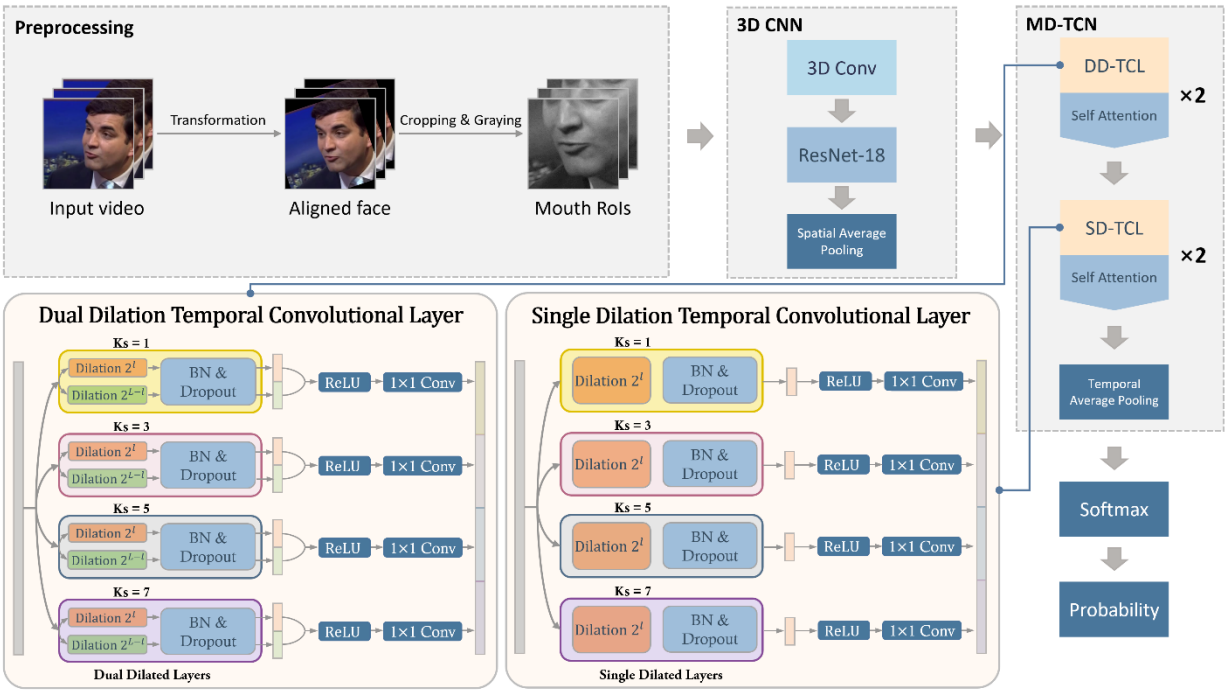

Figure 1: The pipeline of proposed Lip Reading recognition network Self-Attention Multi-Dilation Temporal Convolutional Networks(MD-TCN). "Ks" means kernel size, "BN" means batch normalization, and "Dilation" means dilated factor of the 1D convolution.

3.2. Multi-Dilation TCN

TCN is much better than other sequential models in parallel processing and training speed. However, due to the size limitation of the convolution kernel, the receptive field of TCN is usually limited. In other words, it is difficult to comprehensively consider information with a long interval. The most basic TCN method [15] usually solves this problem by increasing the number of hidden layers and adding dilation layer by layer. Although the top layers may have a broad receptive field, the lowest levels still have a very small receptive field. Furthermore, because of the significant dilation factor of upper layers in TCN, convolutions must be applied at very distant time steps.

Inspired by [21], we adopt a dual dilated convolution (DDC) to replace the traditional dilated convolution. Two convolutions with different dilation factors are combined in DDC (shown as the orange and green blocks in Fig. 1). Smaller levels of the first convolution (orange block in Fig. 1) have a lower dilation factor, which increases exponentially as the number of layers increases. The second

convolution (green block in Fig. 1) starts with a strong dilation factor in the lower levels and gradually decreases as the number of layers grows. Finally, in order to ensure that the output shape is similar to standard TCN, we use a 1*1 convolution layer to transform the shape. Since higher layer don't have the problem of conception field size, we only use DDC in the lower layers of the MD-TCN, while we still use traditional Single Dilation Convolution in the higher layers of the network.

In the meantime, because the size of the conventional TCN convolution kernel is constant, all activations of a certain layer have the same temporal receptive field. As a result, this kind of network generally cannot consider both long-term and short-term information. We expect that the network will be able to capture receptive fields over a range of time scales, allowing short-term and long-term data to be combined for feature encoding. We use a TCN with several convolution kernels to do this. Each temporal convolution in this multi-kernel TCN variation now has many branches, each with a distinct kernel size. Each convolutional layer therefore combines data from many time scales.

In view of the above two points, we finally generate the Multi-Dilation Temporal Convolutional Networks (MD-TCN) to replace the standard TCN (shown in Fig. 1). In this network, we create four temporal branches with kernel sizes of 1, 3, 5, and 7, respectively. And in each branch, we also use dual dilated convolution to replace ordinary dilated convolution. Hence each layer of this MB-TCN has eight branches. After each convolution, we employ Batch Norm layers [24] to speed up training converge, and we apply dropouts [25] using dropping probability of 0.5 for regularization. Meanwhile, as in standard TCN, we also reuse two identical convolutional layer in each MB-TCN to achieve better model effectiveness.

In our experiments, we adopt a total of four convolutional layers structure, because this setup can balance the training speed and accuracy. Also, the number of layers of the hyperparameter DDC in our model is set to 2, which is proved to be the best value in our experiments, and the specific experiments will be shown later in the discussion of hyperparameters.

3.3. Self-Attention

Each lip position in a lip motion model is frequently linked to other positions in the sequence. If each word corresponds to a position, the lips will likewise be in a regular posture. This prompted us to create a method that would allow us to choose the most relevant context for the features we needed to extract. As a result, we propose a temporal attention strategy built in TCN for assigning weights to contextual information at each time step in an adaptable manner. We incorporate a self-attention block after each MB-TCN to account for the autocorrelation of lip-reading sequences.

4. Experiment

4.1. Dataset

Our studies were done using the Lip Reading in the Wild (LRW) [11] dataset, which is the biggest publicly accessible dataset for lipreading individual English words. The LRW dataset has a vocabulary of 500 English words. Each video sequence segment in LRW has a length of 1.16 seconds and was recorded from over 1000 speakers in a BBC show (29 video frames). This dataset contains 538 766 sequences, which are separated into 488 766/25 000/25 000 for training, validation, and testing. Due to the enormous number of speakers and the wide changes in lighting conditions, head positions, and speech speeds, this dataset is also one of the most difficult dataset in Lip Reading.

4.2. Experiment Setup

4.2.1. Preprocessing:

We preprocessed the video similar to the methods introduced in [17]. We first detected face marks and did face alignments for every single video. Then we can easily crop the videos into the size of 96×96 and converted them to grayscale. In order to simulate different lighting and positions between

different videos, we did a bunch of data argumentations such as random horizontal flip, random brightness jitter 20% and random contrast jitter 20%. Finally, to avoid over-fitting to training dataset, we randomly select 1 to 3 frames in a video and randomly delete or copy them. Thus our model can be more powerful to fit different application scenarios.

4.2.2. Pretraining:

Since the easy part of the dataset is often correct even for the simplest model, the accuracy of the model is ultimately determined by the most difficult part of the dataset. We observed that pretrain the whole model on a relatively small sub-dataset is also an effective way to adjust hyperparameters, test model performance and accelerate training. We extract the 50 hardest words based on the current state-of-the-art open-source model [17] for LRW to create the sub-dataset. As a result, we use this pre-training method since it may significantly speed up training.

4.2.3. Training Settings:

The whole model is trained end-to-end, with all weights initialized using the pretrained model, as illustrated in Fig. 1. With a batch size of 32, an initial learning rate of 0.0004, and a weight decay of 1e-4, we train for 90 epochs. To acquire the weights at the best performing point, we measure the accuracy using the validation set provided by LRW. We adopt Adam [22] and cross entropy as optimizer and loss function. The learning rate is gradually reduced from its original value to zero using the cosine scheduling [23].

4.2.4. Implementations:

Our approach is conducted in the PyTorch framework 3.8.10 [27]. We used a 1080Ti GPU for our experiments. The LRW sub-dataset takes roughly 5 hours to train an end-to-end hyperparametric validation model. To train a single model from beginning to end, it takes roughly 5 days. Our network is lighter than other works, and our training time is significantly lowered.

4.2.5. Explorations of Hyperparameter:

On the LRW sub-dataset, we analyze several structural parameters of the MD-TCN model in order to find the best performing one. In particular, we validate the effect of the selection of DDC layers on the results while freezing other hyperparameters, such as the total number of layers and the kernel size. To accelerate training, we only train on LRW sub-dataset for 20 epochs and fix the weight of the front-end (ResNet).

4.3. Results

Table 1
Performance With Different DDC Layers

Number of DDC layer(s)	Test Accuracy (%)
0	72.8
1	72.6
2	73.5
3	73.4
4	73.6

4.3.1. Number of DDC layers:

To determine the best DC-TCN structure, we evaluated the effect of number of dual dilated layers on the LRW sub-dataset, while keeping the values of other hyperparameters constant. As shown in Table 1, it is noticed that the best performance is achieved when the number of DDC layers is 4, i.e., all dilated layers are dual dilated layers. The performance is significantly increased when the number of DDC layer is greater than or equal to two. We note that with more than two layers, increasing DDC layers will barely improve the test accuracy, so we decided to use two DDC layers in the subsequential experiments for a good balance between accuracy and speed.

Table 2
Performance With Use of Self-Attention

Self-Attention Block	Test Accuracy (%)
Yes	73.5
No	73.1

4.3.2. Attention block:

For effectiveness of attention block, we also do an experiment on whether the Self-Attention Block is enabled. Here we do this experiment by using the DDC layers of 2 to ensure consistency with the final experimental hyperparameters. The result shows that when using Self-Attention Block, the training speed slow down but can achieve a better test accuracy. It is worth stating that since the sub-dataset was selected from the most difficult 10% of LRW and we only trained for 20 epochs, the results shown here are not as accurate as they could be.

Table 3
Comparison with Other Methodologies on LRW

Method	Front-end	Back-end	Acc. (%)
LRW [11]	VGG-M	-	61.1
WLAS [5]	VGG-M	LSTM	76.2
ResNet+BLSTM [26]	3D Conv + ResNet34	BLSTM	83.0
End-to-end AVR [28]	3D Conv + ResNet34	BLSTM	83.4
Multi-Grained [29]	ResNet34 + DenseNet3D	Conv-BLSTM	83.3
Multi-scale TCN [17]	3D Conv + ResNet18	MS-TCN	85.3
Face cutout [30]	3D Conv + ResNet18	BGRU	85.0
Multi-modality SR [31]	3D ResNet50	TCN	84.8
3D-ResNet+Bi-GRU [10]	3D SE-ResNet	Bi-GRU	85.5
Ours	3D Conv + ResNet18	MD-TCN	85.7

4.3.3. Performance on LRW:

On the LRW dataset, we compare the performance of our technique versus several baseline methods in Table 3. On the LRW dataset, our technique achieves an accuracy of 85.7 percent, which has a 0.2 percent increase over other similarly structured networks [10].

4.4. Discussion

4.4.1. Effectiveness of LRW Sub-dataset:

We chose the LRW sub-dataset as the training dataset for both hyperparameter tuning and pre-training because of accelerated training. However, we did not investigate whether the LRW sub-dataset can truly reflect the LRW dataset and whether the hyperparameters that perform well in the LRW sub-dataset also perform well in the LRW dataset. Therefore, we extracted the results of the completed training model to investigate how accurate the part of the results corresponding to the sub-dataset.

4.4.2. Optimizability:

Although our model is fast and has very good accuracy, there is still much room for optimization. First, although the LRW sub-dataset is selected from the LRW, the data might be different in distribution, so the network architecture applicable to the sub-dataset may not be optimal on the LRW. Therefore, with sufficient arithmetic power, it is better to select hyperparameters directly on the LRW to guarantee optimal results. Second, in this paper, we improve the accuracy by constructing a dual dilation layer, where the dilation can also be carefully designed like using more than two dilation in one layer. Third, we have taken the data pre-processing approach of adding and deleting frames for data argumentation, but this may affect the speaker's rhythm. We would like to take the approach of directly increasing or decreasing the video length (without changing the frame rate) to simulate different speech rates of the speaker, which has the potential to significantly improve the model robustness.

5. Conclusion

In this work, we propose MD-TCN for isolated word recognition. Our model has a very good performance on LRW by solving the problem of sparse perception field and data autocorrelation. We also adopted a novel method for data argumentation that randomly copy and delete frames to improve model robustness. Finally, we use sub-dataset to select hyperparameters and accelerate training.

6. References

[1] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 17(3):423–435, 2009.

[2] Xiaopeng Hong, Hongxun Yao, Yuqi Wan, and Rong Chen, “A pca based visual dct feature extraction method for lip-reading,” in 2006 International Conference on Intelligent Information Hiding and Multimedia. IEEE, 2006, pp. 321–326.

[3] 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.

[4] 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.

[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] Wand, Michael, Jan Koutník, and Jürgen Schmidhuber. "Lipreading with long short-term memory." In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6115-6119. IEEE, 2016.

[7] Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. "Lipnet: End-to-end sentence-level lipreading." arXiv preprint arXiv:1611.01599 (2016).

[8] Themos Stafylakis, Muhammad Haris Khan, and Georgios Tzimiropoulos, “Pushing the boundaries of audiovisual word recognition using residual networks and lstms,” Computer Vision and Image Understanding, vol. 176, pp. 22–32, 2018.
[9] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic, “Towards practical lipreading with distilled and efficient models,” ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.
[10] Feng, Dalu, Shuang Yang, Shiguang Shan, and Xilin Chen. "Learn an effective lip reading model without pains." arXiv preprint arXiv:2011.07557 (2020).
[11] Chung, Joon Son, and Andrew Zisserman. "Lip reading in the wild." Asian conference on computer vision. Springer, Cham, 2016.
[12] Yang, Shuang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild." In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1-8. IEEE, 2019.
[13] 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.
[14] 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.
[15] Lea, Colin, Rene Vidal, Austin Reiter, and Gregory D. Hager. "Temporal convolutional networks: A unified approach to action segmentation." In European Conference on Computer Vision, pp. 47-54. Springer, Cham, 2016.
[16] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018.
[17] Martinez, Brais, Pingchuan Ma, Stavros Petridis, and Maja Pantic. "Lipreading using temporal convolutional networks." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6319-6323. IEEE, 2020.
[18] Wan, Renzhuo, Shuping Mei, Jun Wang, Min Liu, and Fan Yang. "Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting." Electronics 8, no. 8 (2019): 876.
[19] Dai, Rui, Luca Minciullo, Lorenzo Garattoni, Gianpiero Francesca, and François Bremond. "Self-attention temporal convolutional network for long-term daily living activity detection." In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-7. IEEE, 2019.
[20] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.
[21] Li, Shi-Jie, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. "Ms-tcn++: Multi-stage temporal convolutional network for action segmentation." IEEE transactions on pattern analysis and machine intelligence (2020).
[22] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
[23] Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." arXiv preprint arXiv:1608.03983 (2016).
[24] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In International conference on machine learning, pp. 448-456. PMLR, 2015.
[25] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15, no. 1 (2014): 1929-1958.
[26] Stafylakis, Themos, and Georgios Tzimiropoulos. "Combining residual networks with LSTMs for lipreading." arXiv preprint arXiv:1703.04105 (2017).

[27] Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen et al. "Pytorch: An imperative style, high-performance deep learning library." Advances in neural information processing systems 32 (2019).
[28] Petridis, Stavros, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. "End-to-end audiovisual speech recognition." In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 6548-6552. IEEE, 2018.
[29] Wang, Chenhao. "Multi-grained spatio-temporal modeling for lip-reading." arXiv preprint arXiv:1908.11618 (2019).
[30] Zhang, Yuanhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition." In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 356-363. IEEE, 2020.
[31] Xu, Bo, Cheng Lu, Yandong Guo, and Jacob Wang. "Discriminative multi-modality speech recognition." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14433-14442. 2020.