Multi-Task Learning for Audio-Visual Active Speaker Detection

Yuan-Hang Zhang[1,2*] Jingyun Xiao[1,2*] Shuang Yang[1,2] Shiguang Shan[1,2,3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China
{zhangyuanhang15,xiaojingyun15}@mails.ucas.ac.cn
{shuang.yang,sgshan}@ict.ac.cn

Abstract

This report describes the approach underlying our submission to the active speaker detection task (task B-2) of ActivityNet Challenge 2019. We introduce a new audio-visual model which builds upon a 3D-ResNet18 visual model pretrained for lipreading and a VGG-M acoustic model pretrained for audio-to-video synchronization. The model is trained with two losses in a multi-task learning fashion: a contrastive loss to enforce matching between audio and video features for active speakers, and a regular cross-entropy loss to obtain speaker / non-speaker labels. This model obtains 84.0% *mAP on the validation set of AVA-ActiveSpeaker. Experimental results showcase the pretrained embeddings' abilities to transfer across tasks and data formats, as well as the advantage of the proposed multi-task learning strategy.*

1. Introduction

Research on audio-visual analysis of speech contents in videos has received increasing attention for its academic and practical value. Some important topics include visual speech recognition or lipreading, and audio-visual speech separation and enhancement. However, although there has been a great amount of work in these fields, few works have explored active speaker detection, which aims to determine which, if any, of the visible people in a video is speaking at any given time. This is a necessary and important pre-processing step for visual speech recognition and other downstream applications such as video conferencing. Previous models for this task were often trained and evaluated on videos recorded in constrained environments, which hinders the construction of robust active speaker detection models. In this paper, we introduce a new deep audiovisual model for the task and evaluate on the recently released AVA-ActiveSpeaker dataset[4] to investigate its potentials in real-life, in-the-wild settings.

Another open question is whether state-of-the-art models trained for the lipreading task can be transferred directly to the speaker detection task, since both tasks rely heavily on the fine-grained analysis of lip motion. At present, while commonly used training data for both tasks are branded as "in-the-wild" datasets, they differ significantly in terms of domain (e.g. news, documentaries, and drama vs. movies) and quality (e.g. the presence of tiny faces, occlusion, and extreme poses). We aim to answer this question using the new AVA-ActiveSpeaker dataset.

2. Methods

The network is depicted in Fig. 1, composed of a visual subnetwork and an audio subnetwork which extract per-frame features, a two-layer bidirectional GRU that operates on concatenated features, and a final classifier to obtain speaker/non-speaker labels. During training and evaluation, the inputs to the network are sequences of T frames and the corresponding audio representations (which are of variable lengths, due to difference in frame rates), and the network outputs T speaker / non-speaker labels for each frame. We next expand on the design of each module and explain our training strategy.

2.1. Visual Features

The visual features are extracted with the 3D-ResNet18 model first proposed in [6], which has been widely adopted for lipreading, and also used for audio-visual speech enhancement. It begins with a spatio-temporal convolution layer with kernel size 5 in the temporal dimension, which effectively models the short-term dynamics of visual speech, and then progressively reduces spatial dimensionality with an 18-layer residual network (ResNet-18). The

*Equal contribution.

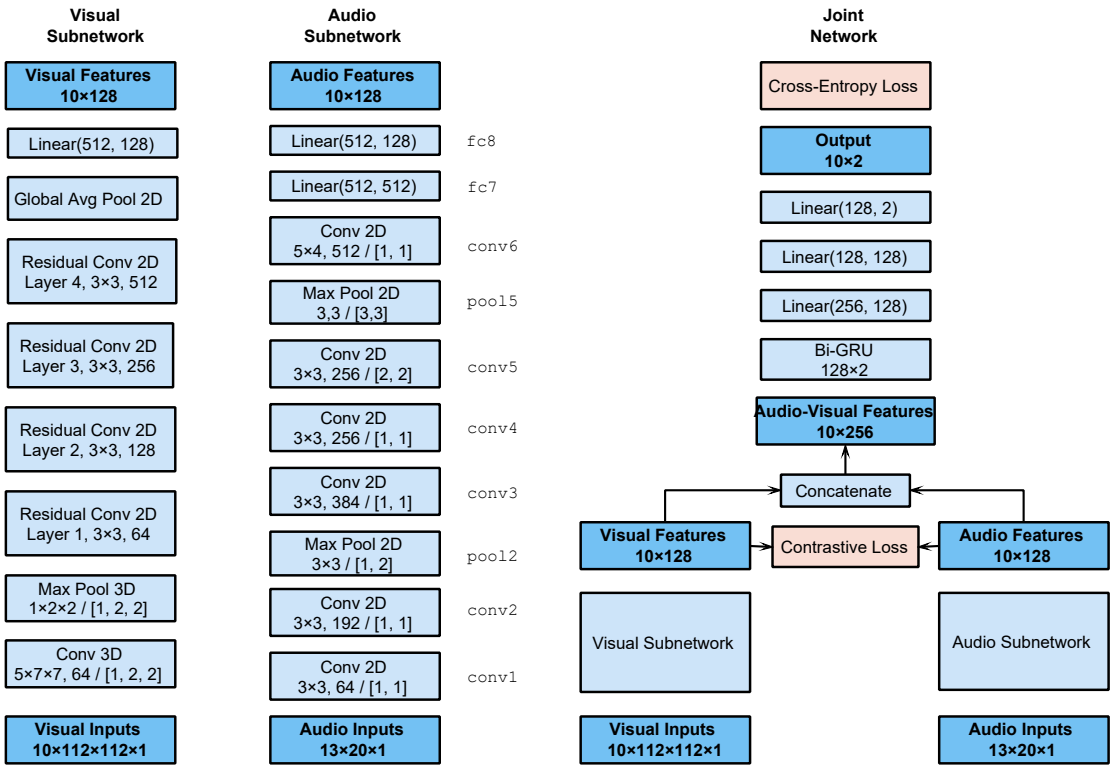

Figure 1. **Network architecture**. We build upon 3D-ResNet and VGG-M subnetworks. For a window with T frames, we first extract per-frame audio and visual features, and then feed the concatenated features to a 2-layer bidirectional Gated Recurrent Unit (GRU) to obtain final framewise predictions.

average-pooled feature is transformed into an 128-dim visual embedding with the final fully-connected layer. We initialize the subnetwork using weights pre-trained on the LRW, LRS, and MV-LRS datasets [1][1].

We did not perform temporal up / downsampling for videos recorded at different frame rates during feature extraction. This is because some speech/non-speech segments can be very short (e.g. lasting 3-5 frames) and downsampling can negatively impact framewise accuracy. Moreover, we believe a model that generalizes well should be robust to mild frame rate differences, which can be viewed as natural differences in speech rates between individual speakers.

2.2. Audio Features

The audio features are obtained using a modified VGG-M network which ingests 13-dim MFCC features as input. The features are extracted using a 25ms analysis window with a stride of 10ms, yielding 100 audio frames every second. The network we use is pre-trained with the improved two-stream SyncNet architecture [2, 3] for audio-to-video synchronization[2]. For each video frame, we extract audio embeddings by calculating the corresponding audio time and passing the surrounding 20 audio frames to the network. To reduce the dimensionality of the output features, we retrain the final `fc8` layer to obtain 128-dim embeddings.

2.3. Feature Fusion

We concatenate the per-frame 128-dim visual features and the 128-dim audio features generated by the visual subnetwork and the audio subnetwork. The joint features are fed into a two-layer bidirectional GRU with 256 cells for temporal modeling within the input window, which is complementary to the spatio-temporal convolution in the visual subnetwork. Finally, the outputs of the last hidden layer are fed to a classifier with 3 fully-connected layers, which predicts for every time-step a binary label indicating whether the specified face is the person speaking or not.

[1] https://github.com/afourast/deep_lip_reading

[2] https://github.com/joonson/syncnet_python

2.4. Multi-Task Learning

Inspired by the design of SyncNet for audio-to-video synchronization, we apply an extra contrastive loss on the visual and audio features to enforce a match between the two modalities. Since the audio inputs and visual inputs are synchronized, if the ground truth label is `SPEAKING_AUDIBLE`, it means the speaker in the visual input is speaking and the audio input matches the visual input, and the Euclidean distance between the two features should therefore be minimized. If the ground truth label is `NOT_SPEAKING` or `SPEAKING_NOT_AUDIBLE`, the speaker in the visual input is not speaking and the visual input does not match the audio input. We expect that this form of contrastive learning would help the network better capture information that is shared between audio and video, and therefore better utilize the multi-modal cues.

Training is performed in two stages: we first load the weights pre-trained on large-scale datasets, and initialize the last fully-connected layer in each subnetwork with only the contrastive loss while keeping other layers frozen. After convergence, we add the fusion network and train the joint model end-to-end in a multi-task learning fashion: at each time step, the loss is a combination of the above contrastive loss and the routine cross-entropy loss between the predicted and ground truth labels. The overall loss is the aggregated loss over all time steps:

$$\mathcal{L} = \sum_{t=1}^{T} (\mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{contrast}}), \tag{1}$$

where λ is a hyperparameter. In our experiments, we use $\lambda = 1$.

3. Experiments

Preprocessing. We first run face detection with the multi-view face detector provided with the SeetaFaceEngine2 toolkit [5], and compare them with the annotations provided with the AVA-ActiveSpeaker dataset. The detected bounding boxes were then smoothed with median filtering. This yields much tighter and more accurate face crops[3], which we scale properly to match the pre-training settings. Finally, we extract a central 112×112 crop from each face and use it as the input to the visual subnetwork.

Implementation details. We implement our model with PyTorch and train the model on the training set with the Adam optimizer, using an initial learning rate of 1e-4. We use batches of 40 windows with window length $T = 10$, and data augmentation in the form of horizontal flipping. The images are converted to grayscale and normalized with respect to overall mean and variance. We train the model for around 20 epochs, reducing learning rate whenever error on the validation set plateaus.

The windows we use for training are sampled densely from all tracks at a stride of 1. Since there are more frames labeled as non-speaker ($682,404$ frames labeled `SPEAKING_AUDIBLE`, in contrast to $1,969,134$ marked `NOT_SPEAKING` and $24,776$ marked `SPEAKING_NOT_AUDIBLE` in the training set), we balance training by randomly choosing between windows with more than a half of the frames labeled as active speaker and windows completely labeled as non-active speaker.

Results. Results of our model on the validation set are shown in Table 1. The predicted speech-active scores have been smoothed using a median filter of kernel size 11 within each track. In practice, we found that this yields about 1% performance boost in terms of mAP.

From the table we can see that the pre-trained visual embeddings displays strong performance even when only visual information is used, possibly due to the use of deeper architectures and pre-training. This confirms that the pre-trained embeddings can be transferred across tasks with simple finetuning, depsite the differences in terms of content and quality. Notably, our full model yields about 4.8% performance gain over the ablation model trained without the contrastive loss, which demonstrates the effectiveness of the proposed multi-task learning strategy.

Table 1. Mean average precision (mAP) on the validation and test sets. We do not report ablation results on the test set due to submission timeout policies.

Method	**Visual**		**Audio-Visual**	
	Val	Test	Val	Test
Baseline [4]	/	0.711	/	0.821
Ours (w/o contrastive loss)	0.757	/	0.792	/
Ours	/	/	0.840	0.835

4. Acknowledgements

This work was supported, in part, by the Google Cloud Platform Credits for AVA Challenge participants.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: A comparison of models and an online application. *Proc. Interspeech 2018*, pages 3514–3518, 2018. 2

[3] We found some inaccurate bounding boxes through IoU thresholding and a CNN face/non-face classifier, which were either not centered on the desired subject or did not include the crucial lip region. We corrected these annotations manually, and will make this information available to foster reproducibility.

[2] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2

[3] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019. 2

[4] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava-activespeaker: An audio-visual dataset for active speaker detection. *arXiv preprint arXiv:1901.01342*, 2019. 1, 3

[5] SeetaTech. Seetafaceengine2. https://github.com/seetaface/SeetaFaceEngine2, 2018. 3

[6] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. pages 3652–3656, 2017. 1