NUS-HLT Report for ActivityNet Challenge 2021 AVA (Speaker)

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou and Haizhou Li

National University of Singapore

ruijie.tao@u.nus.edu

Abstract

Active speaker detection (ASD) seeks to detect who is speaking in a visual scene of one or more speakers. The successful ASD depends on accurate interpretation of short-term and long-term audio and visual information, as well as audio-visual interaction. Unlike the prior work where systems make decision instantaneously using short-term features, we propose a novel framework, named TalkNet, that makes decision by taking both short-term and long-term features into consideration. TalkNet consists of audio and visual temporal encoders for feature representation, audio-visual cross-attention mechanism for inter-modality interaction, and a self-attention mechanism to capture long-term speaking evidence. The experiments demonstrate that TalkNet achieves 3.5% and 3.0% improvement over the state-of-the-art systems on the AVA-ActiveSpeaker validation and test dataset, respectively. We will release the codes, the models and data logs.

1. Introduction

As the short-term audio and visual features represent the salient cues for ASD, most of the existing studies are focused on segment-level information, e.g., a video segment of 200 to 600 ms. A better way to capture the long-term temporal context is to encode the history of audio or video frame sequence. In this report, we study an audio-visual ASD framework, denoted as TalkNet. We make the following contributions.

- We propose a feature representation network to capture the long-term temporal context from audio and visual cues;
- We propose a backend classifier network that employs audio-visual cross-attention, and self-attention to learn the audio-visual inter-modality interaction;
- We propose an effective audio augmentation technique to improve the noise-robustness of the model.

2. TalkNet

TalkNet is an end-to-end pipeline that takes the entire cropped face video and corresponding audio as input, and decide if the person is speaking in each video frame. It consists of a feature representation frontend, and a speaker detection backend classifier, as illustrated in Figure 1. The frontend contains an audio temporal encoder and a video temporal encoder. They encode the frame-based input audio and video signals into the time sequence of audio and video embeddings, that represent temporal context. The backend classifier consists of an inter-modality cross-attention mechanism to dynamically align audio and visual content, and a self-attention mechanism to observe speaking activities from the temporal context at the utterance level.

2.1. Visual Temporal Encoder

The visual temporal encoder consists of the visual frontend and the visual temporal network. The visual frontend consists of a 3D convolutional layer (3D Conv) followed by a ResNet18 block [1] to encode the video frame stream into a sequence of frame-based embedding. The visual temporal network consists of a video temporal convolutional block (V-TCN), which has five residual connected rectified linear unit (ReLU), batch normalization (BN) and depth-wise separable convolutional layers (DS Conv1D) [2], followed by a Conv1D layer. It aims to represent the temporal content in a long-term visual spatio-temporal structure. We seek to encode the visual stream into a sequence of visual embeddings F_v that have the same time resolution.

2.2. Audio Temporal Encoder

The audio temporal encoder seeks to learn an audio content representation from the temporal dynamics. It is a 2D ResNet34 network with squeeze-and-excitation (SE) module introduced in [3]. An audio frame is first represented by a vector of Mel-frequency cepstral coefficients (MFCCs). The audio temporal encoder takes the sequence of audio frames as the input, generate the sequence of audio embeddings F_a as the output. The ResNet34 are designed with dilated convolutions such that the time resolution of audio embeddings F_a matches that of the visual embeddings F_v to facilitate subsequent attention mechanism.

2.3. Audio-visual Cross-Attention and Self-Attention

The core part of the cross-attention network is the cross-attention layer, in which one modality applies the query from another modality as the target. The attention layer is followed by the feed-forward layer. Residual connection and layer normalization are also applied after these two layers to generate the whole cross-modal attention network. The outputs are concatenated together along the temporal direction. A self-attention network is applied after the cross-attention network to model the audio-visual utterance-level temporal information. This network is similar to the cross-attention network except that the query, key and value in the attention layer all come from the joint audio-visual feature F_{av}.

2.4. Loss Function

We finally apply a fully connected layer followed by a softmax operation to project the output of the self-attention network to an ASD label sequence. We view ASD as a frame-level classification task instead of a short instance-level classification task. The predicted label sequence is compared with the ground truth label sequence by binary cross-entropy loss.

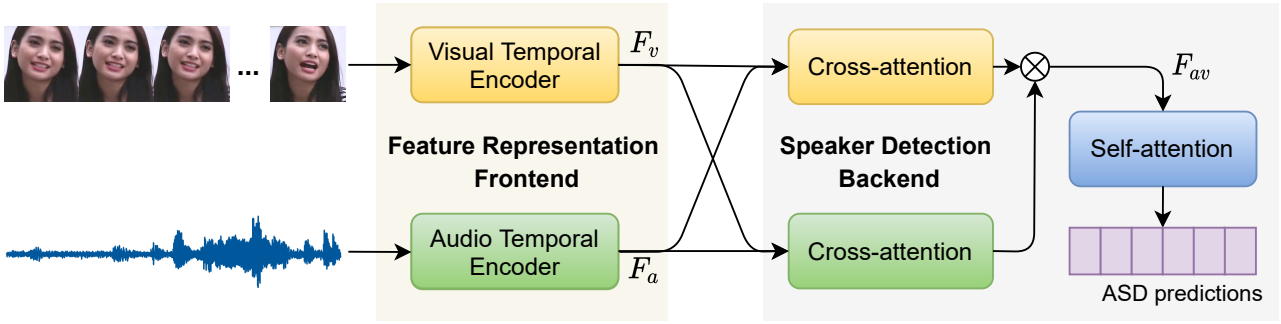

Figure 1: *An overview of our TalkNet, which consists of visual and audio temporal encoders followed by cross-attention and self-attention for ASD prediction. It predicts the entire score sequence from the entire input video without splitting.*

2.5. Audio Augmentation with Negative Sampling

we use one video as the input data during training, and then we randomly select the audio track from another video in the same batch as the noise to perform audio augmentation. Such augmented data effectively have the same label, e.g., active speaker or inactive speaker, as the original sound track. This approach involves the in-domain noise and interference speakers from the training set itself. It does not require data outside the training set.

3. Experiments

We build the TalkNet using the PyTorch library with the Adam optimizer. The initial learning rate is 10^{-4}, and we decrease it by 5% for every epoch. The dimension of MFCC is 13. All the faces are reshaped into 112×112. We set the dimensions of the audio and visual feature as 128. Both cross-attention and self-attention network contain one transformer layer with eight attention heads. We randomly flip, rotate and crop the original images to perform visual augmentation. Finally, we evaluate the performance using the official tool from ActivityNet. There is no any pre-train model and we train only about 7 hours with 20 epochs to get the best result in AVA Validation set from scratch by using one Quadro RTX 6000 GPU with 22G memory.

4. Results

We summarize the results on the AVA-ActiveSpeaker validation dataset in Table 1. We observe that TalkNet achieves 92.3% mAP and outperforms the best competitive system, i.e., MAAS-TAN [4], by 3.5% on the validation set.

Table 1: *Comparison with the state-of-the-art on the AVA-ActiveSpeaker validation set in terms of mean average precision (mAP).*

Method	mAP (%)
Roth et al. [4, 5]	79.2
Zhang et al. [6]	84.0
MAAS-LAN [4]	85.1
Alcazar et al. [7]	87.1
Chung et al [8]	87.8
MAAS-TAN [4]	88.8
TalkNet (proposed)	**92.3**

We obtain the evaluation results of the 'secret' test set in Table 2. Our 90.8% mAP also outperforms the best prior work by 3.0%, cf., Chung et al. [8].

TalkNet only uses the AVA-ActiveSpeaker training set to train the single face videos from scratch without any additional post-processing. We believe that pre-training [6, 8] and other advanced techniques [4, 7] will further improve TalkNet, which is beyond the scope of this work.

Table 2: *Comparison with the state-of-the-art on the AVA-ActiveSpeaker test set in terms of mAP.*

Method	mAP (%)
Roth et al. [5]	82.1
Zhang et al. [6]	83.5
Alcazar et al. [7]	86.7
Chung et al. [8]	87.8
TalkNet (proposed)	**90.8**

5. References

[1] T. Afouras, J. S. Chung, and A. Zisserman, "The Conversation: deep audio-visual speech enhancement," *Proc. Interspeech*, pp. 3244–3248, 2018.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2018.

[3] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech 2020*, 2020.

[4] J. León-Alcázar, F. C. Heilbron, A. Thabet, and B. Ghanem, "Maas: Multi-modal assignation for active speaker detection," *arXiv preprint arXiv:2101.03682*, 2021.

[5] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "AVA active speaker: An audio-visual dataset for active speaker detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*. IEEE, 2020, pp. 4492–4496.

[6] Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan, "Multi-task learning for audio-visual active speaker detection," 2019.

[7] J. L. Alcázar, F. Caba, L. Mai, F. Perazzi, J.-Y. Lee, P. Arbeláez, and B. Ghanem, "Active speakers in context," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 465–12 474.

[8] J. S. Chung, "Naver at activitynet challenge 2019–task b active speaker detection (ava)," *arXiv preprint arXiv:1906.10555*, 2019.