LoCoNet: *Lo*ng-Short *Co*ntext *Net*work for Active Speaker Detection

Supplementary Material

1. Comparison of results of LoCoNet and TalkNet

We show more comparisons and results of LoCoNet and TalkNet [3] on videos from AVA-ActiveSpeaker [2](Fig.1), Ego4D [1](Fig.2) and the Internet(Fig.3). The videos depict several complicated scenes including (1) talking and chewing, (2) egocentric view with head motion, and (3) poker game with narration. Results show that LoCoNet better differentiates active and inactive speakers in these complex scenes with multi-people conversations.

2. Visualizations of the results of LoCoNet

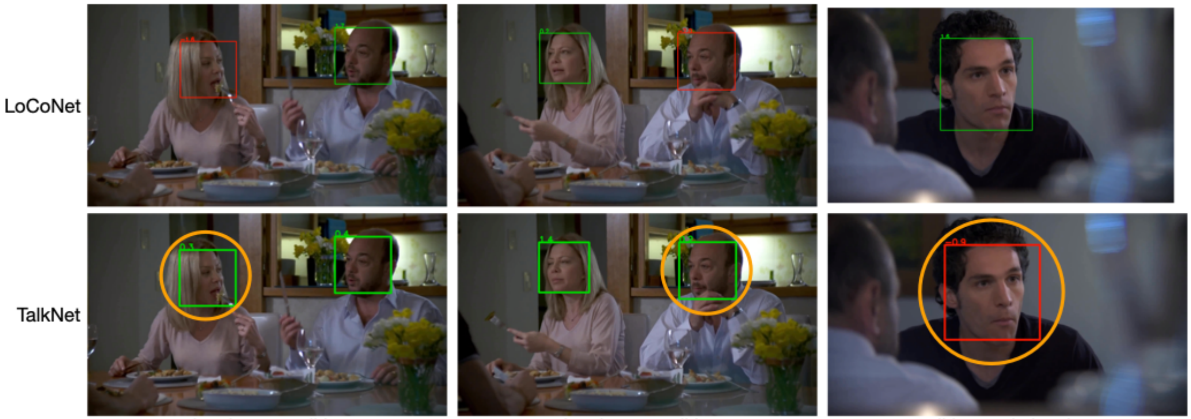
We generate the prediction results of LoCoNet on the videos from AVA-ActiveSpeaker, Ego4D and the Internet, and show the visualizations of the results by mapping them to the original videos. The visualizations are attached as mp4 files in the supplementary material. In the videos, the green boxes denote active speakers. The red boxes denote inactive speakers. The numbers above the boxes are logits where those larger than 0 are predicted as active speakers, and those smaller than 0 are predicted as inactive speakers.

References

[1] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

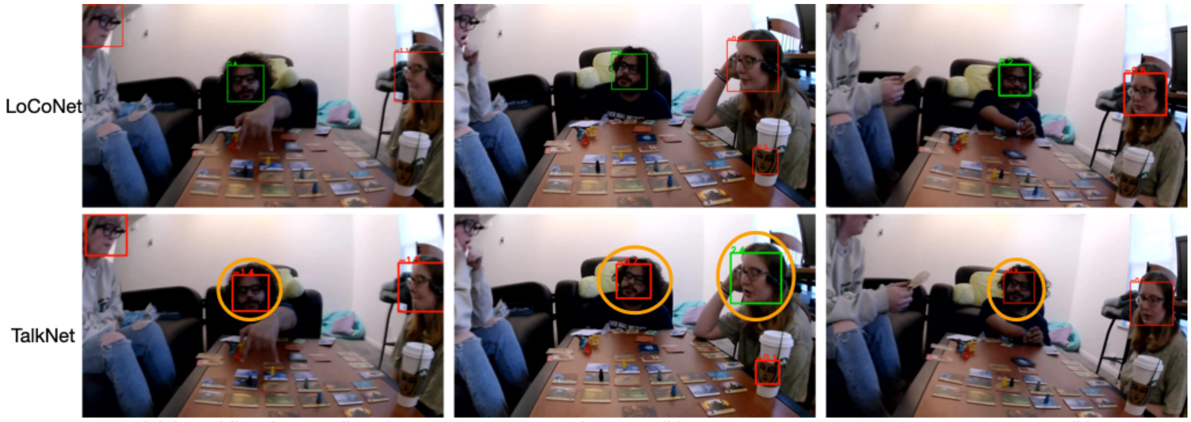
[2] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020. 1

[3] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 1, 2



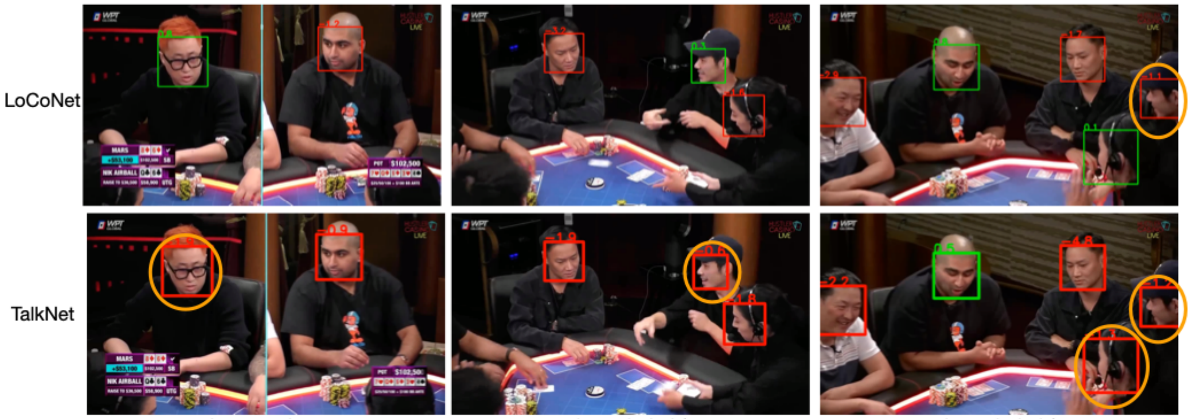
chewing not talking | chewing not talking | talking and chewing

Figure 1. **Results of LoCoNet and TalkNet on a video from AVA-ActiveSpeaker validation set.** The green boxes denote predicted active speaker. The red boxes denote predicted inactive speaker. The orange boxes denote false predictions. This video shows a conversation on the dinner table, and the mouth movements of the speakers might be talking, chewing, or both.



explaining while other two listening | take turn talking | camera-wearer talking

Figure 2. **Results of LoCoNet and TalkNet [3] on a video from Ego4D dataset.** This egocentric video shows a group of friends playing board games. The head motion of the camera-wearer and rapid shift in conversations make the inference more challenging.



short words | partly occluded face | partly occluded face

Figure 3. **Results of LoCoNet and TalkNet [3] on a video from the Internet.** This video shows a livestream of poker game with real-time narration. The results show that predicting speaking activities of a speaker with partly occluded face remains a difficult task.