Mutual Alignment between Audiovisual Features for End-to-End Audiovisual Speech Recognition

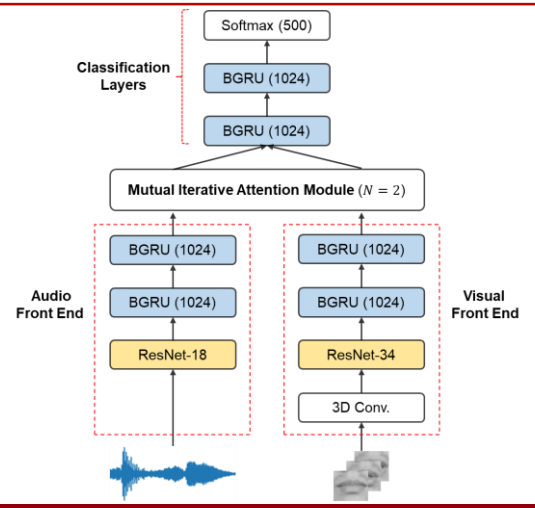
Hong Liu, *Yawei Wang*, *Bing Yang*

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University

{hongliu, Wongyawei}@pku.edu.cn, bingyang@sz.pku.edu.cn

Introduction

- **AVSR**: Using complementarity and redundancy between audio and visual speech to improve the accuracy and robustness of speech recognition systems under noise conditions.
- **Challenges**: Most AVSR systems assume that the audio and visual features are synchronized and are concatenated directly [1]. Both AV Align [2] and AliNN [3] apply additive attention within seq-to-seq architecture, but they overly rely on one modality and fail to align two modality features precisely under some noise conditions.
- **Major contribution**: we adopt a mutual feature alignment method [4] where the features from one modality can be utilized as the guide for aligning the features of the other modality iteratively to make full use of cross modality information in the process of alignment.



End-to-End Framework

Mutual Iterative Attention

- Positional Encoding

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{input}}),\ PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{input}})$$

- Scaled Dot-Product Attention

$$\mathrm{Att}_i(Q,S) = \mathrm{softmax}(\frac{QW_i^Q(SW_i^K)^\top}{\sqrt{d_k}})SW_i^V, i = 1,...,k$$

- Multi-Head Attention

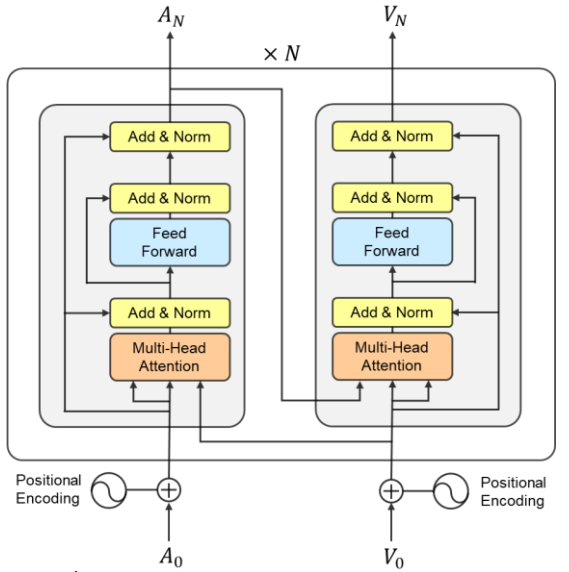
$$\mathrm{MultiHead}(Q,S) = [\mathrm{Att}_1(Q,S),...,\mathrm{Att}_k(Q,S)]W^O$$

- Feed Forward Network

$$\mathrm{FFN}(X) = \max(0, XW_1 + b_1)W_2 + b_2$$

- Mutual Attention

$$A' = \mathrm{FFN}(\mathrm{MultiHead}(V,A)),\ V' = \mathrm{FFN}(\mathrm{MultiHead}(A',V))$$



- Mutual Iterative Attention

$$A_1 = \mathrm{FFN}(\mathrm{MultiHead}(V_0,A_0)),\ V_1 = \mathrm{FFN}(\mathrm{MultiHead}(A_1,V_0))$$

.....

$$A_N = \mathrm{FFN}(\mathrm{MultiHead}(V_{N-1},A_{N-1})),\ V_N = \mathrm{FFN}(\mathrm{MultiHead}(A_N,V_{N-1}))$$

- National Natural Science Foundation of China
- National Natural Science Foundation of Shenzhen

Acknowledgement

Experimental Results

Table. Recognition performance in word classification rate [%] of various models on LRW dataset at different SNR levels. AV_baseline is the feature concatenation based method where the audio and visual sequences are assumed to be aligned frame-by-frame.

Model	Word Classification Rate (%)							
	clean	20dB	15dB	10dB	5dB	0dB	-5dB	AVG
audio-only	96.74	96.68	96.48	95.85	94.07	88.07	68.90	90.97
visual-only	77.24	77.24	77.24	77.24	77.24	77.24	77.24	77.24
AV_baseline	97.42	97.38	97.36	97.12	96.49	94.22	87.17	95.31
AV_MIA(Ours)	**97.55**	**97.54**	**97.48**	**97.27**	**96.82**	**94.92**	**89.32**	**95.84**

Conclusion

- In this work, a mutual feature alignment method is proposed to address the asynchronization issue in audiovisual speech recognition.
- We introduce Mutual Iterative Attention mechanism to align the audio and visual features by performing mutual attention over the two modalities iteratively.
- Our proposed method outperforms the feature concatenation based AVSR system over all noisy conditions.

1. Petridis S, Stafylakis T, Ma P, et al. End-to-end audiovisual speech recognition. *In ICASSP*, 2018: 6548-6552.
2. Sterpu G, Saam C, Harte N. Attention-based audio-visual fusion for robust automatic speech recognition. *In ICMI*, 2018: 111-115.
3. Tao F, Busso C. Aligning audiovisual features for audiovisual speech recognition. *In ICME*, 2018: 1-6.
4. Liu F, Liu Y, Ren X, et al. Aligning visual regions and textual concepts for semantic-grounded image representations. *In NIPS*, 2019: 6850-6860.

References