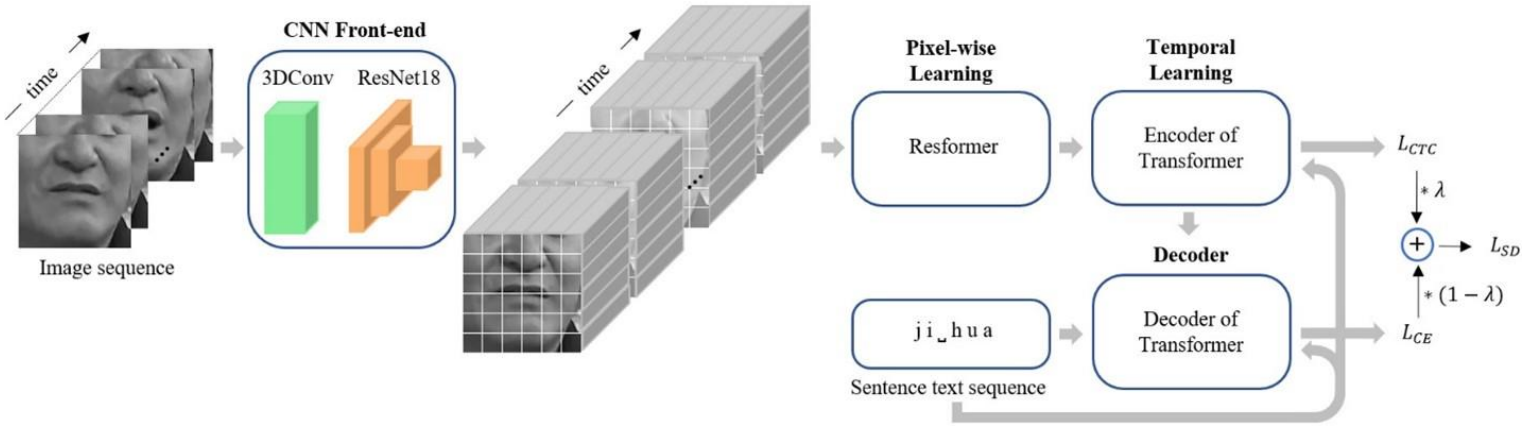
Fine-grained Sequence-to-Sequence Lip Reading Based on Self-Attention and Self-Distillation

Junxiao XUE, Shibo HUANG, Huawei SONG, Lei SHI

Frontiers of Computer Science, DOI: 10.1007/s11704-023-2230-x

Problems & Ideas

- Problems of lip reading approaches:
  - Lip reading is a fine-grained video analysis that requires both the local information and the overall spatial information of the sequence.
  - Most existing approaches capture local spatial information with CNN and temporal information with RNN generally.
- Ideas: we propose a self-attention and self-distillation method for fine-grained Seq2Seq lip reading.



Framework of Seq2Seq lip reading based on self-attention and self-distillation

Main Contributions

- Contributions:
  - Resformer module to extract fine-grained features from lip images;
  - Self-distilling method to further improve lip-reading accuracy;
  - The overall structure of the lip-reading model in this paper is a Seq2Seq structure. We conducted experiments on GRID, LRW and LRW-1000 datasets, and achieved significant results.

Datasets	Methods	CER↓	WER↓
GRID	$\lambda = 0$	0.579%	2.357%
	$\lambda = 0.1$	0.552%	2.236%
	$\lambda = 0.3$	0.541%	2.163%
	$\lambda = 0.5$	0.544%	2.201%
	+ SE Blocks ($\lambda = 0.3$)	**0.458%**	1.846%
	+ Resformer ($\lambda = 0.3$)	0.459%	**1.774%**
LRW	$\lambda = 0$	17.442%	22.696%
	$\lambda = 0.1$	15.478%	20.276%
	$\lambda = 0.3$	12.400%	16.444%
	$\lambda = 0.5$	12.783%	17.200%
	+ SE Blocks ($\lambda = 0.3$)	11.807%	15.780%
	+ Resformer ($\lambda = 0.3$)	**11.023%**	**14.752%**
LRW-1000	$\lambda = 0$	53.918%	65.048%
	$\lambda = 0.1$	49.829%	59.413%
	$\lambda = 0.3$	47.812%	57.650%
	$\lambda = 0.5$	51.130%	61.929%
	+ SE Blocks ($\lambda = 0.3$)	45.944%	55.771%
	+ Resformer ($\lambda = 0.3$)	**44.381%**	**54.603%**

The results of ablation experiments on GRID, LRW and LRW-1000 datasets