SAT-Net: Self-Attention and Temporal Fusion for Facial Action Unit Detection

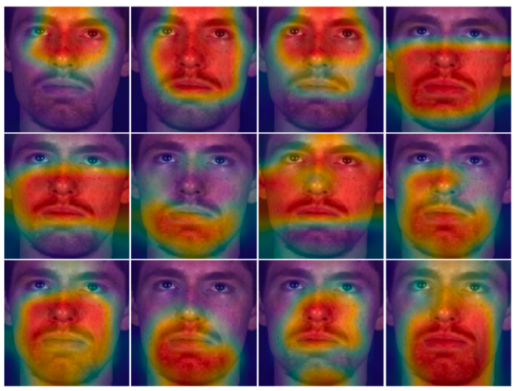
Zhihua Li, Zheng Zhang, Lijun Yin
Binghamton University
BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

➢ Intro to facial action units

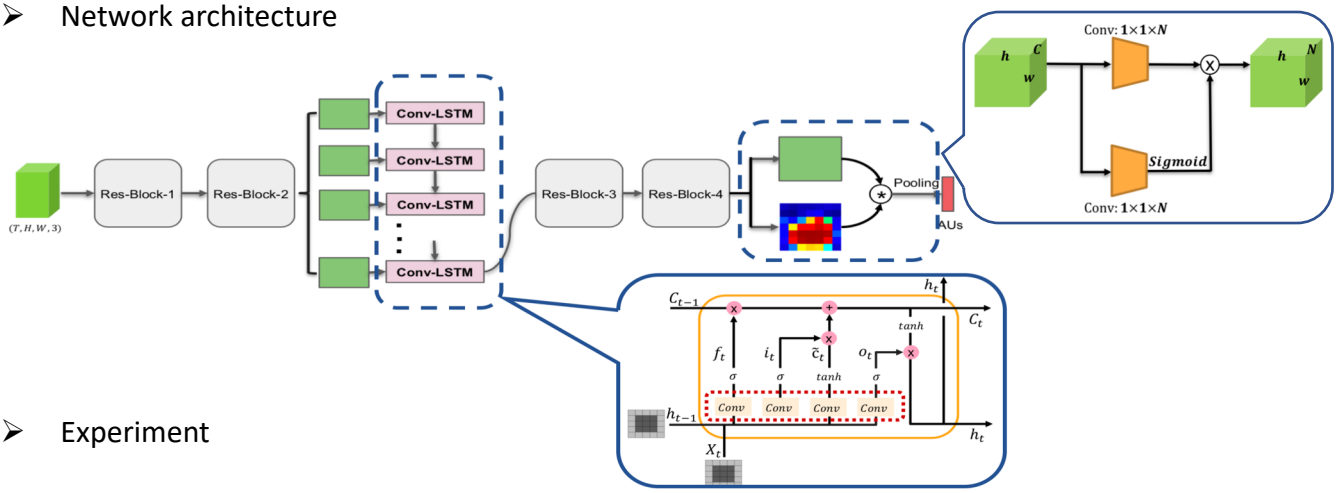
Facial action units (FAUs) are facial muscle actions at certain facial locations defined by Facial Action Coding System (FACS) Ekman, R. (1997)

AU index	Au Name
1	Inner Brow Raiser
2	Outer Brow Raiser
4	Brow Lowerer
6	Cheek Raiser
7	Lid Tightener
10	Upper Lip Raiser
12	Lip Corner Puller
14	Dimpler
15	Lip Corner Depressor
17	Chin Raiser
23	Lip Tightener
24	Lip Pressor

➢ Attention maps



➢ Network architecture



➢ Experiment

AU	JPML	DRML	CNN-LSTM	EAC	JAA	LP	$AR_{ConvLSTM}$	SRERL	ResNet	T-Net	SA-Net	SAT-Net
1	32.6	36.4	31.4	39.0	47.2	43.4	48.0	46.9	50.8	45.9	52.0	[54.1]
2	25.6	41.8	31.1	35.2	44.0	38.0	43.2	45.3	45.4	43.9	45.1	[49.5]
4	37.4	43.0	71.4	48.6	54.9	54.2	53.1	55.6	56.2	55.8	[60.0]	58.3
6	42.3	55.0	63.3	76.1	77.5	77.1	76.9	77.1	77.1	76.5	[78.0]	77.7
7	50.5	67.0	77.1	72.9	74.6	76.7	78.4	[78.4]	76.6	76.8	76.9	77.7
10	72.2	66.3	45.0	81.9	[84.0]	83.8	82.8	83.5	82.3	82.6	83.8	83.6
12	74.1	65.8	82.6	86.2	86.9	87.2	[87.9]	87.6	86.7	86.8	87.3	86.5
14	65.7	54.1	[72.9]	58.8	61.9	63.3	67.7	60.6	57.2	59.5	61.0	63.2
15	38.1	36.7	33.2	37.5	43.6	45.3	45.6	52.2	49.3	[53.0]	49.5	49.1
17	40.0	48.0	53.9	59.1	60.3	60.5	63.4	[63.9]	60.5	62.8	61.0	61.8
23	30.4	31.7	38.6	35.9	42.7	48.1	47.9	47.1	48.1	[50.0]	47.5	48.7
24	42.3	30.0	37.0	35.8	41.9	54.2	[56.4]	53.3	50.0	48.5	47.6	49.3
Avg	45.9	48.3	53.2	55.9	60.0	61.0	62.6	62.9	61.7	61.8	62.5	[63.3]

Table 1: Performance on BP4D

AU	LSVM	APL	DRML	EAC	JAA	$AR_{ConvLSTM}$	SRERL	ResNet	SA-Net	T-Net	SAT-Net
1	10.8	11.4	17.3	41.5	43.7	26.9	[45.7]	29.8	32.3	36.6	41.2
2	10.0	12.0	17.7	26.4	46.2	24.4	[47.8]	29.3	33.1	32.5	33.1
4	21.8	30.1	37.4	[66.4]	56.0	58.6	59.6	56.6	62.3	64.9	63.0
6	15.7	12.4	29.0	50.7	41.4	49.7	47.1	[57.3]	52.2	53.1	56.4
9	11.5	10.1	10.7	[80.5]	44.7	34.2	[45.6]	35.5	33.3	35.8	43.0
12	70.4	65.9	37.7	[89.3]	69.6	71.3	73.5	71.8	71.2	74	73.1
25	12.0	21.4	38.5	[88.9]	88.3	83.4	84.3	84.6	84.0	82.2	82.9
26	22.1	26.9	20.1	15.6	58.4	51.4	43.6	55.2	59.6	55.7	[60.6]
Avg	21.8	23.8	26.7	48.5	56.0	50.0	55.9	52.5	53.5	54.3	[56.7]

Table 2: Performance on DISFA

➢ Conclusion

- Our network utilized the **least training parameters** but achieves the state-of-the-art performance.
- Different from handcraft attention mechanism, we developed an AU label supervised **self-learned attention module** to enable the network to learn to pay more attention to different facial areas for the corresponding AUs
- We have also proposed to use **Conv-LSTM module** to fuse the temporal information into AU detection problems and proved to be feasible with temporal information as a supplement in facial action unit detection