endAffectNet: Self-Attention based Networks for Predicting Affective Responses from Movies

hi Phuong Thao*, Balamurali B.T.*, Dorien Herremans* and Gemma Roig+ (thiphuongthao_ha@mymail.sutd.edu.sg)
gapore University of Technology and Design (SUTD), +Goethe University Frankfurt, Germany

SINGAP
TECHN

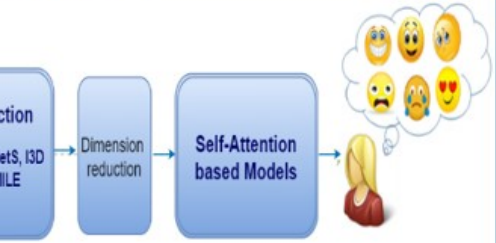
otivation

clips) still remains a challenge

oked in a person is hard for computers.

ive responses of viewers from movies use low-
pproaches

r Approach



roach for predicting emotions from movies.

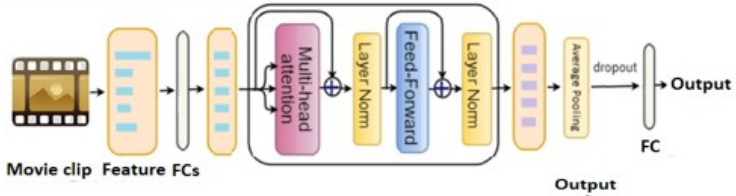
t valence and arousal separately

io

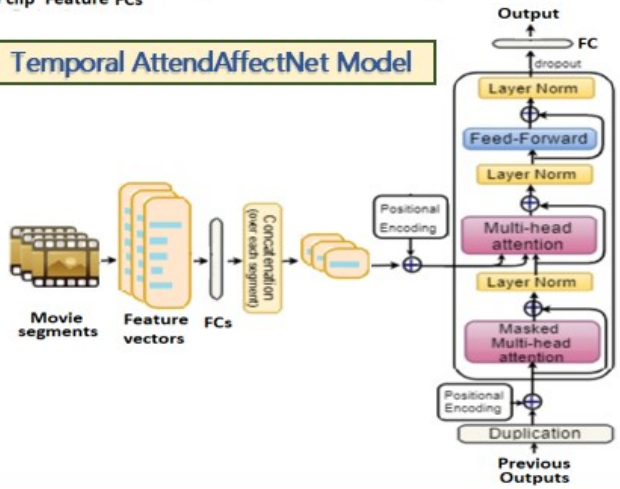
ntion mechanism (Vaswani et al, 2017)

Proposed Models

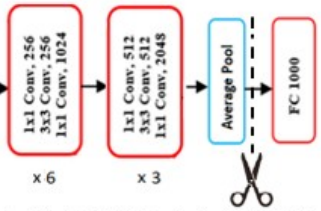
- Figure 6: Feature AttendAffectNet Model



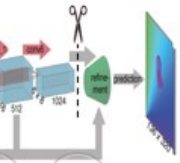
- Figure 7: Temporal AttendAffectNet Model



Feature Extraction



sing ResNet-50 [He et al., CVPR 2016]



t al., ICCV, 2015]

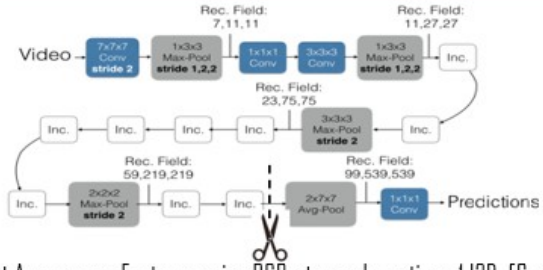


Figure 3: Extract Appearance Features using RGB-stream Inception-v1 I3D [Carreira et al., CVPR 2017]

- Audio
  - OpenSMILE: "emobase2010.conf" (INTERSPEECH 2010): window size = 320ms, hop size = 40ms
  - VGGish:

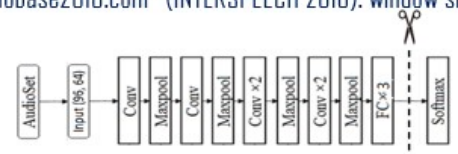


Figure 5: VGGish[Hershey et al., ICASSP, 2017]

Experimental Resul

TABLE 1: ACCURACY OF THE PROPOSED MODELS ON THE EXTENDED COGNIMUSE DATASET.

Models	Arousal MSE	Arousal PCC	Valence MSE	Valence PCC
Feature AAN (only video)	0.152	0.518	0.204	0.483
Feature AAN (only audio)	0.125	0.621	0.185	0.543
Feature AAN (video and audio)	0.124	0.630	0.178	0.572
Temporal AAN (only video)	0.178	0.457	0.267	0.232
Temporal AAN (only audio)	0.162	0.472	0.247	0.254
Temporal AAN (video and audio)	0.153	0.551	0.238	0.319
Sivaprasad et al. [23]	**0.08**	**0.84**	**0.21**	**0.50**

TABLE 2: ACCU
STA

Models
Feature AAN (
Feature ANN (
Feature ANN (
Temporal ANN
Temporal ANN
Temporal ANN
Liu et al. [56]
Chen et al. [55]
Yi et al. [22]
Yi et al. [41]

Visualization

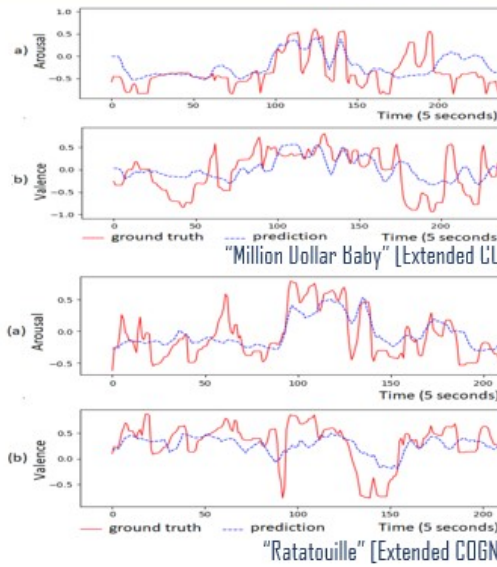


Figure 8: Visualization of the ground-tru

Conclusion

- Feature AttendAffectNet outperforms the Temporal AttendAf
- Audio-based model provides a higher performance than vide
- Model combining all features (video, audio) reaches the high