ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

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

Ha Thi Phuong Thao[*], Balamurali B.T.[*], Dorien Herremans[*], Gemma Roig[+]

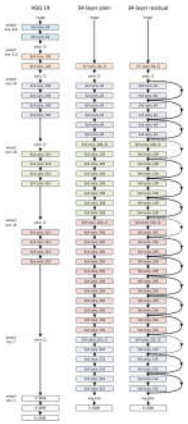
[*]Singapore University of Technology and Design (SUTD)
[+]Goethe University Frankfurt

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

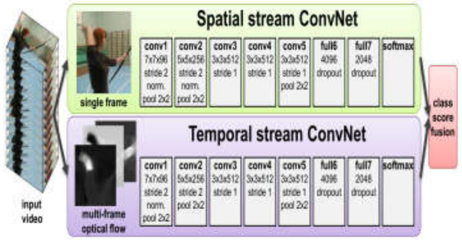
GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Motivation

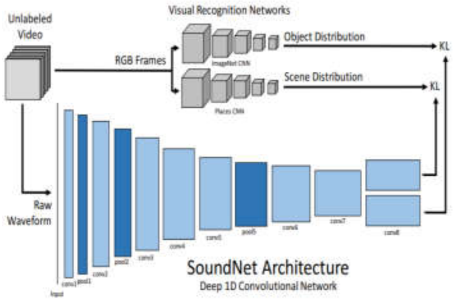
- Breakthroughs in deep CNNs for image classification, action recognition and sound classification,….



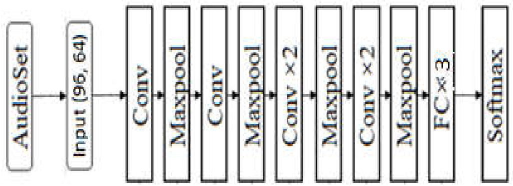
ResNet
[He et al., CVPR 2016]



Two-stream CNN
[Simonyan et al., NIPS, 2014]



SoundNet
[Aytar et al., NIPS, 2016]

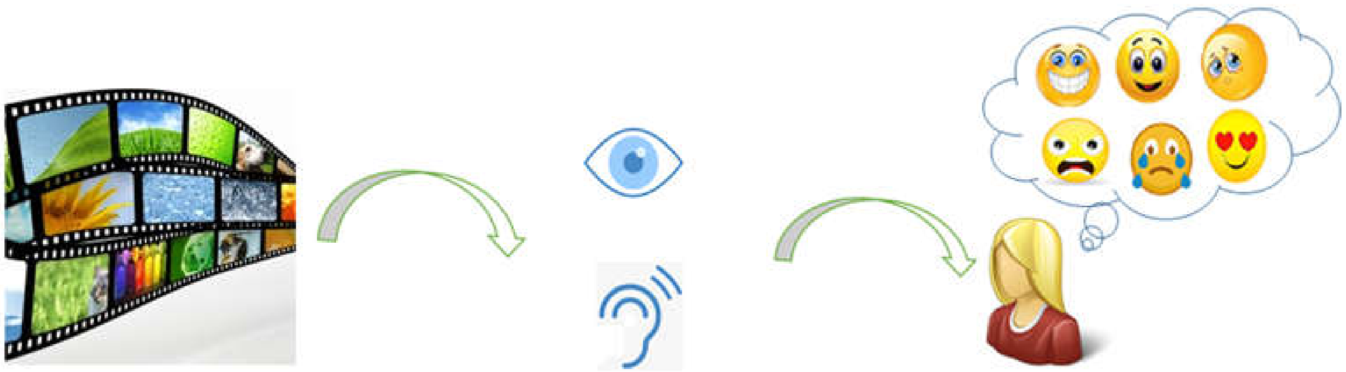


VGGish
[Hershey et al., ICASSP, 2017]

Processing videos (movies, music clips) still remains a challenge

Motivation

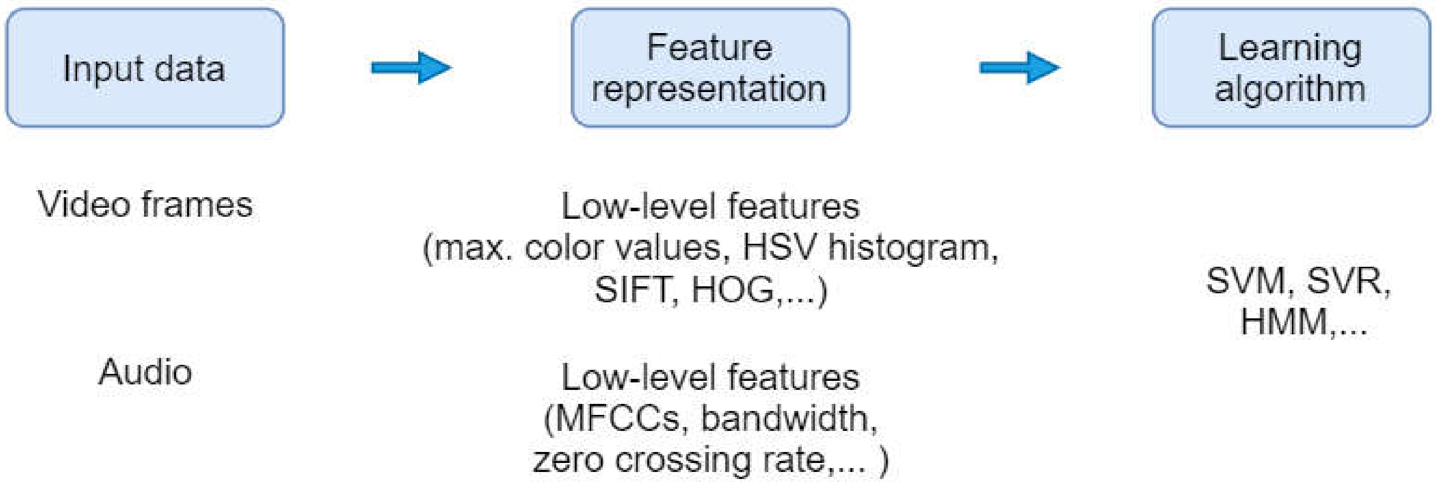
- ❑ Predict what kind of emotion evoked in a person => Hard for computers



Movies

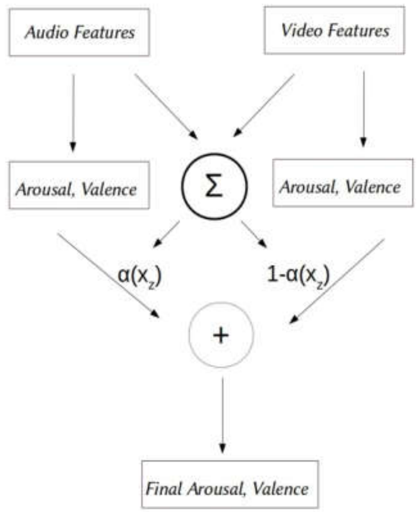
Motivation

- Many studies on predicting affective responses of viewers from movies

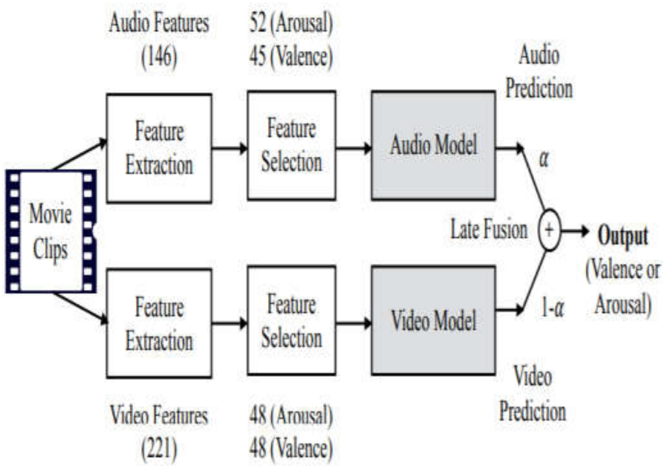


Related Work

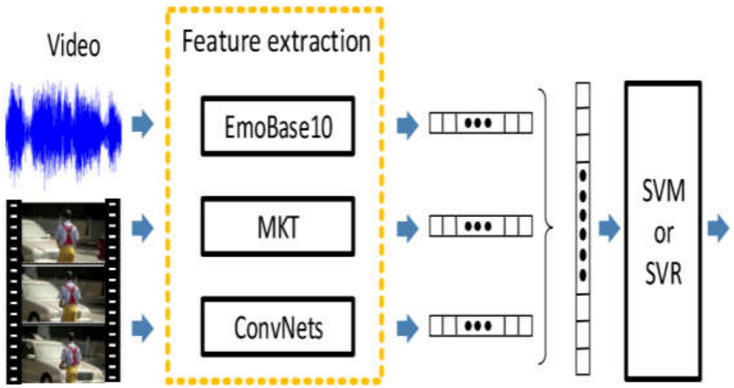
- Focus on fusion techniques, do not explicitly consider the relation among multiple modalities



[Goyal et al., 2016]



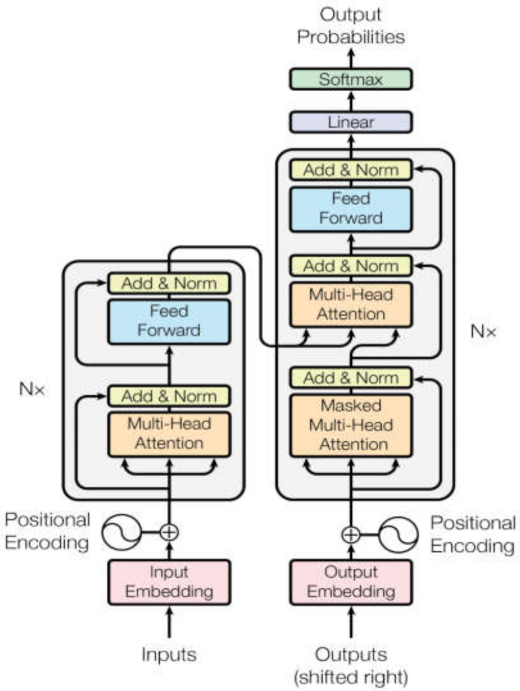
[Sivaprasad et al, 2019]



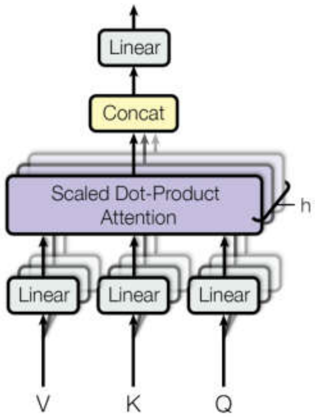
[Yi et al, 2019]

Related Work

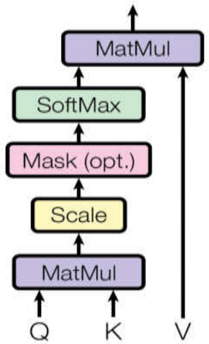
- ❑ Transformer network (Vaswani, NIPS, 2017) => self-attention mechanism could capture temporal/spatial dependencies of input sequence



Transformer Network



Multi-head attention



$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right)$$

Scaled dot-product attention

Positional encoding:

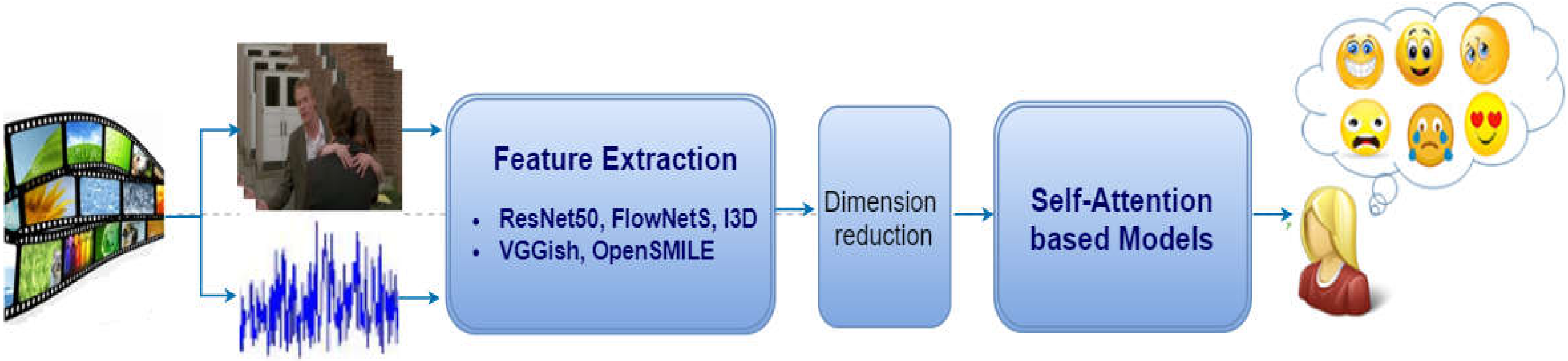
$$\overrightarrow{PE}_{pos}^{(i)} = \begin{cases} \sin(\omega_k . pos) & \text{if } i = 2k \\ \cos(\omega_k . pos) & \text{if } i = 2k+1, \end{cases}$$

$$\text{where } \omega_k = \frac{1}{10000^{2k/d}}$$

$$i = 0, \ldots, d.$$

d : encoding dimension

Approach

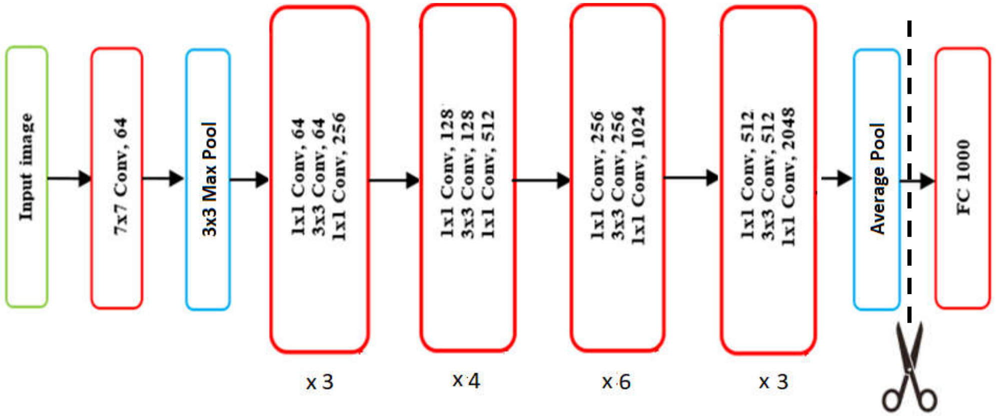


Idea: **multimodal approach to predict valence and arousal separately and directly**

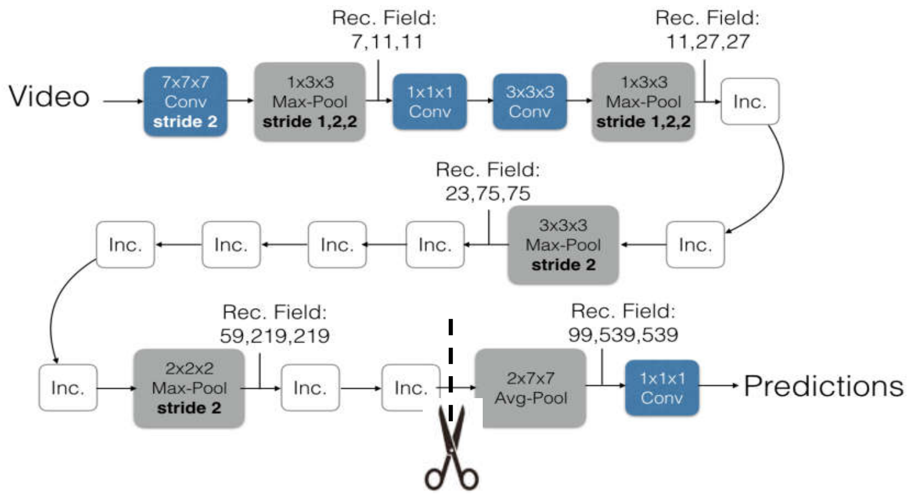
- Features: appearance, motion, audio features
- Models: based on the self-attention mechanism (Vaswani et al, 2017)

Feature Extraction: Appearance

- ❑ Movie excerpts of different length => FFMPEG tool: obtain T frames from each movie excerpt
- ❑ ResNet-50 (pre-trained on ImageNet): static appearance of objects from still frames
- ❑ RGB-stream I3D (pre-trained on Kinetics): spatio-temporal features (appearance and temporal relation)



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



RGB-stream Inception-v1 I3D [Carreira et al., CVPR 2017]

- Each frame: 2048- feature vector
- Element-wise averaging over all frames/each movie excerpt

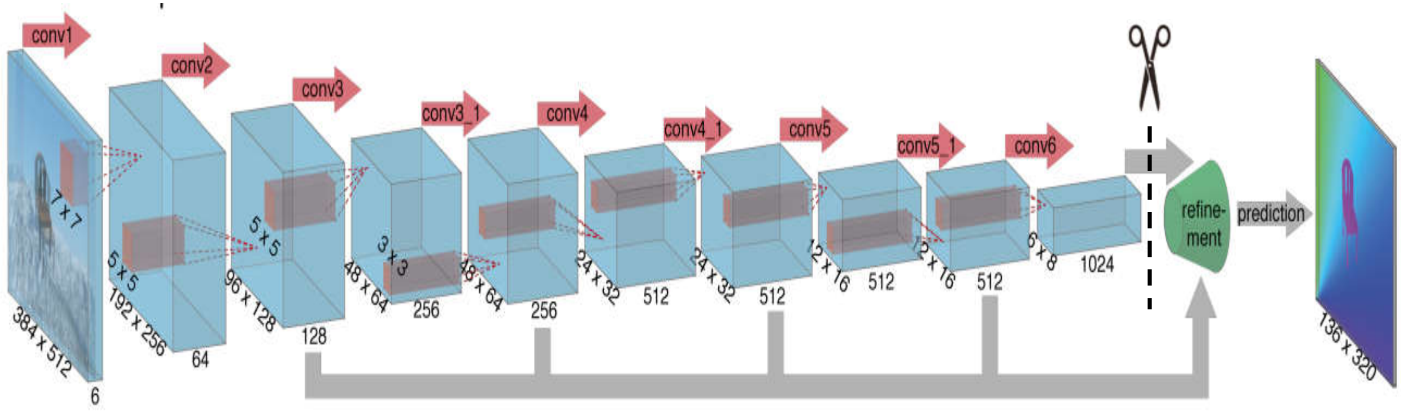
=> 2048-feature vector/movie excerpt

- Remove all after the last Inception module (i.e. "mixed_5_c" layer)
- Input: T frames (C x T x H x W) => Output: $1024 \times \frac{T}{8} \times \frac{H}{32} \times \frac{W}{32}$
- Average Pooling: kernel size of $\frac{T}{8} \times \frac{H}{32} \times \frac{W}{32}$

=> 1024- feature vector/each movie excerpt

Feature Extraction: Motion

- ❑ Optical flow estimation: Expensive!!!
- ❑ FlowNet Simple: pre-trained on Flying Chairs dataset (Dosovitskiy et al., ICCV, 2015)



FlowNet Simple
[Dosovitskiy et al., ICCV, 2015]

- Each pair of consecutive frames: 1024-feature vector
- Element-wise averaging over all pairs of frames/each movie excerpt

  => 1024-feature vector/movie excerpt

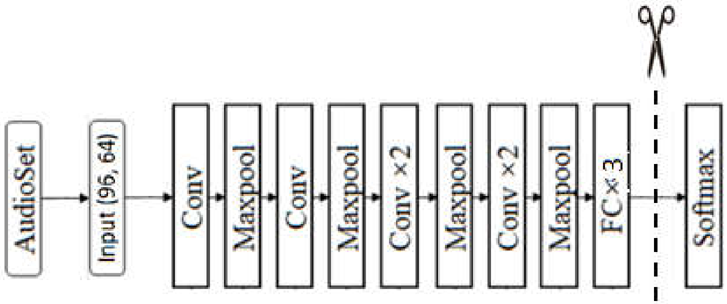
Feature Extraction: Audio

- OpenSMILE:

  - "emobase2010.conf" (INTERSPEECH 2010 paralinguistics challenge)
  - Window size = 320ms, hop size = 40ms => 1,582 features.
  - Element-wise averaging over all 320-ms windows/each movie excerpt

    => 1,582-feature vector/movie excerpt

- VGGish: pre-trained on AudioSet (sound classification)

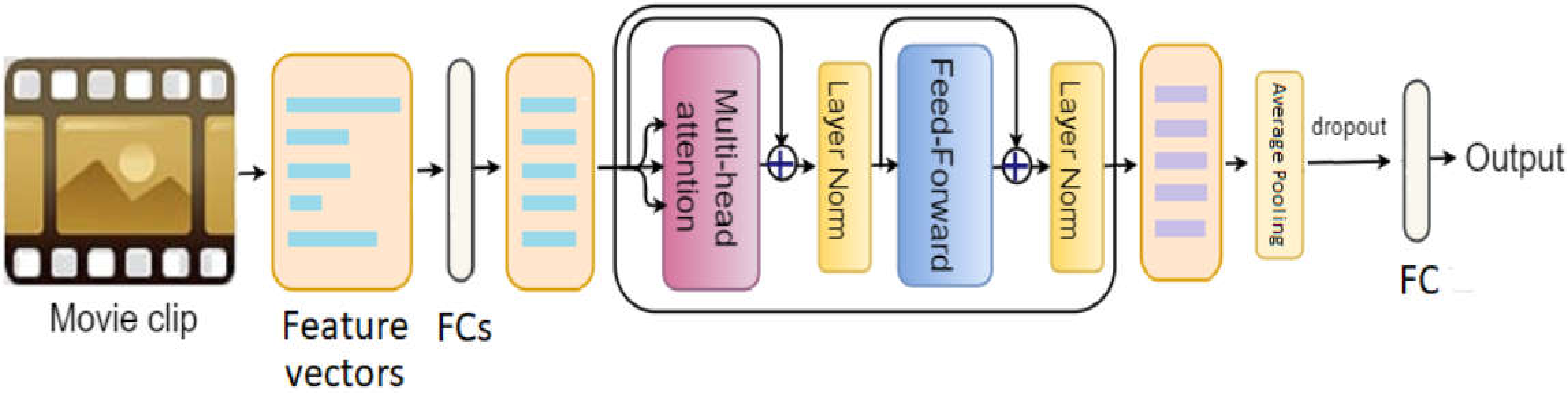


VGGish [Hershey et al., ICASSP, 2017]

  - Each 0.96s => 128 features
  - Element-wise averaging over all 0.96-s audio segments/each movie excerpt

  => 128-feature vector/movie excerpt

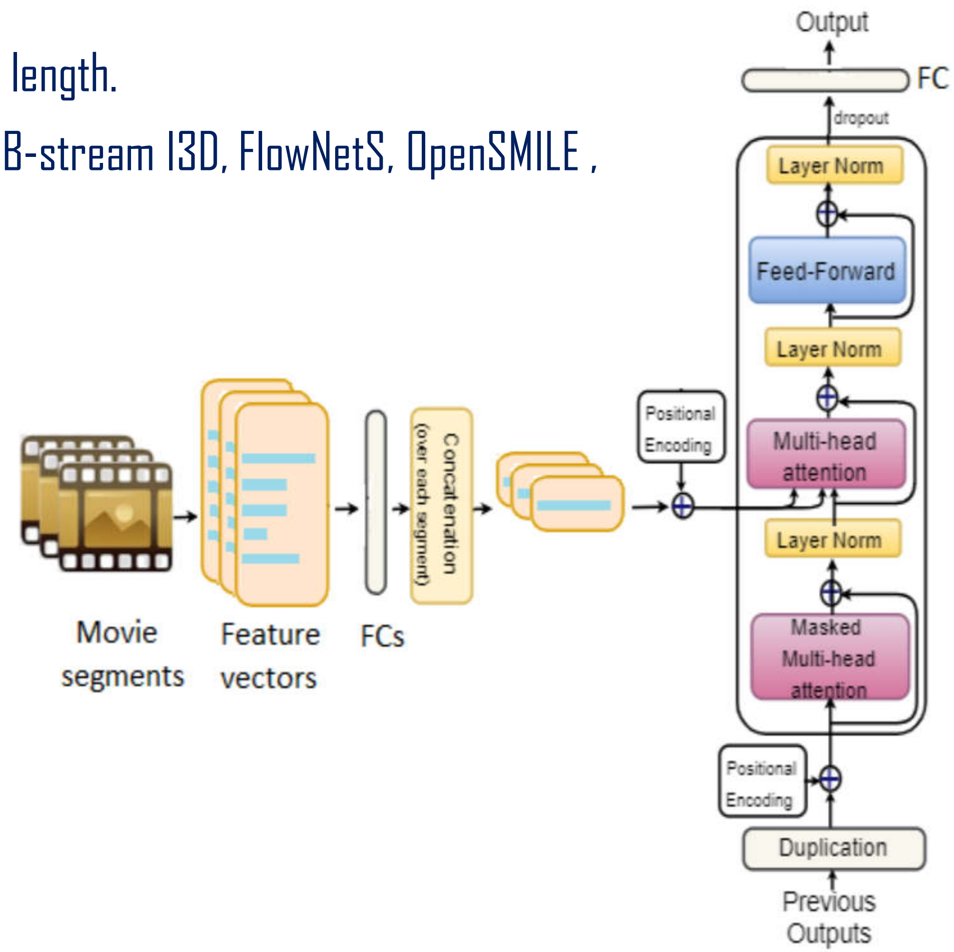
Proposed Model 1: Feature AttendAffectNet



- 5 feature vectors: ResNet-50 (2048), RGB-stream I3D (1024), FlowNetS (1024), OpenSMILE (1582), VGGish (128)

Proposed Model 2: Temporal AttendAffectNet

- Split each original movie clip into segments of the same length.
- Each movie segment: 5 feature vectors (ResNet-50, RGB-stream I3D, FlowNetS, OpenSMILE , VGGish)



Results: Extended COGNIMUSE and Global EIMT16

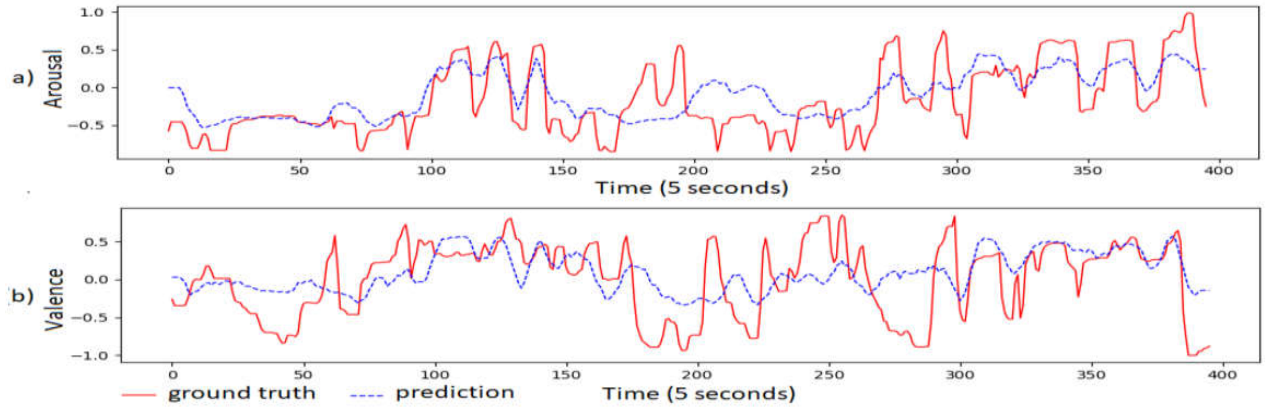
ACCURACY OF THE PROPOSED MODELS ON THE EXTENDED COGNIMUSE DATASET.

	Arousal		Valence	
Models	MSE	PCC	MSE	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**

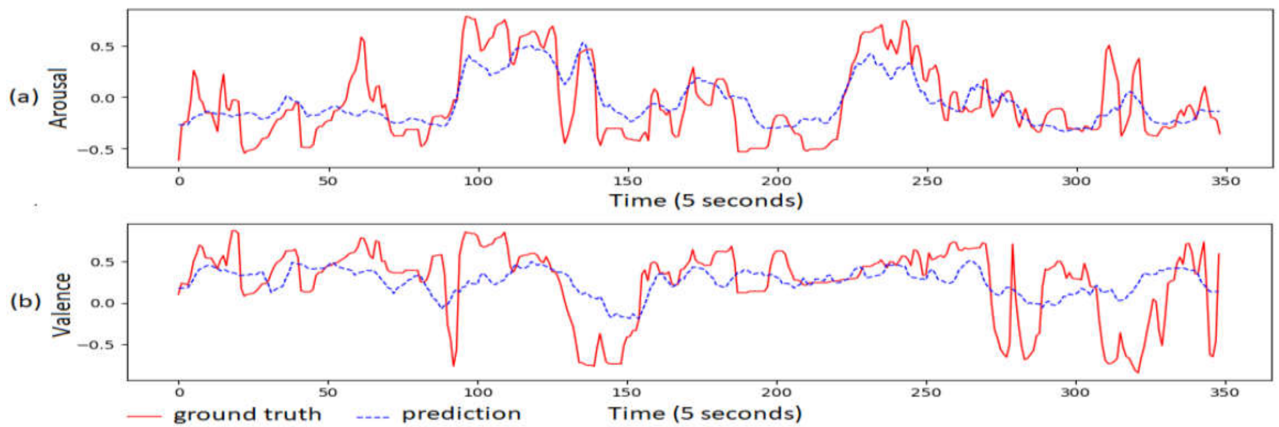
ACCURACY OF THE PROPOSED MODELS IN COMPARISON WITH STATE-OF-THE-ART ON THE GLOBAL EIMT16.

	Arousal		Valence	
Models	MSE	PCC	MSE	PCC
Feature AAN (only video)	0.933	0.350	0.764	0.342
Feature ANN (only audio)	1.111	0.397	0.209	0.327
Feature ANN (video and audio)	**0.742**	**0.503**	**0.185**	**0.467**
Temporal ANN (only video)	1.182	0.151	0.256	0.190
Temporal ANN (only audio)	1.159	0.185	0.225	0.285
Temporal ANN (video and audio)	**0.854**	**0.210**	**0.218**	**0.415**
Liu et al. [56]	1.182	0.212	0.236	0.379
Chen et al. [55]	1.479	0.467	0.201	0.419
Yi et al. [22]	1.173	0.446	0.198	0.399
Yi et al. [41]	**0.542**	**0.522**	**0.193**	**0.468**

Visualization



“Million Dollar Baby” [Extended COGNIMUSE]



“Ratatouille” [Extended COGNIMUSE]

Summary

- Use pre-trained deep neural networks and the OpenSMILE toolkit to extract features from audio and video
- Compare different ways to integrate the extracted features using the self-attention based networks.
- The AttendAffectNet models trained on audio features outperforms those on video features
- Model combining all features (video, audio) reaches the highest performance

Acknowledgement

This work is supported by:

- MOE Tier 2 grant no. MOE2018-T2-2-161
- SRG ISTD 2017 129

Source code:

https://github.com/ivyha010/AttendAffectNet