Deep Regression of Social Signals in Dyadic Scenarios

Author: Ítalo Vidal Lucero
Director: Dr. Sergio Escalera
Co-directors: Dr. Julio Jacques & Cristina Palmero

Agenda

01 Introduction

Introduction • Objectives.

02 State of the art methods

Multi-modal Databases • Feature Extraction • Multi-modal methods

03 Proposal

Independent Feature Extraction • Joint Optimization

04 Experiments

Features • Multi-modal benchmark • Qualitative study • Ablation study.

05 Conclusions

Conclusions • Future work.

Agenda

01 Introduction
Introduction • Objectives.

02 State of the art methods
Multi-modal Databases • Feature Extraction • Multi-modal methods

03 Proposal
Independent Feature Extraction • Joint Optimization

04 Experiments
Features • Multi-modal benchmark • Qualitative study • Ablation study.

05 Conclusions
Conclusions • Future work.

Introduction

- Human communication.
- Deep neural networks as a tool for emotion analysis.
- Affective Computing for empathic machines.
- Scope: - Recognition of six emotional states: happy, angry, neutral, frustrated, excited and sad.
  - Multiple **modalities**: text, audio and video.
  - At utterance level (unit of speech).

Objectives

01 Investigate the public state-of-the-art databases used in dyadic scenarios.

02 Perform an utterance feature extraction pipeline from raw modalities in the study.

03 Reproduce a state-of-the-art multi-modal system in terms of performance.

04 Analyze and propose a state-of-the-art feature representation at the utterance level for text modality.

05 Analyze and propose a joint optimization of modalities for feature representation at the utterance level.

06 Perform an ablation study to compare modalities importance at the utterance level.

Agenda

01 Introduction

Introduction • Objectives.

02 State of the art methods

Multi-modal Databases • Feature Extraction • Multi-modal methods

03 Proposal

Independent Feature Extraction • Joint Optimization

04 Experiments

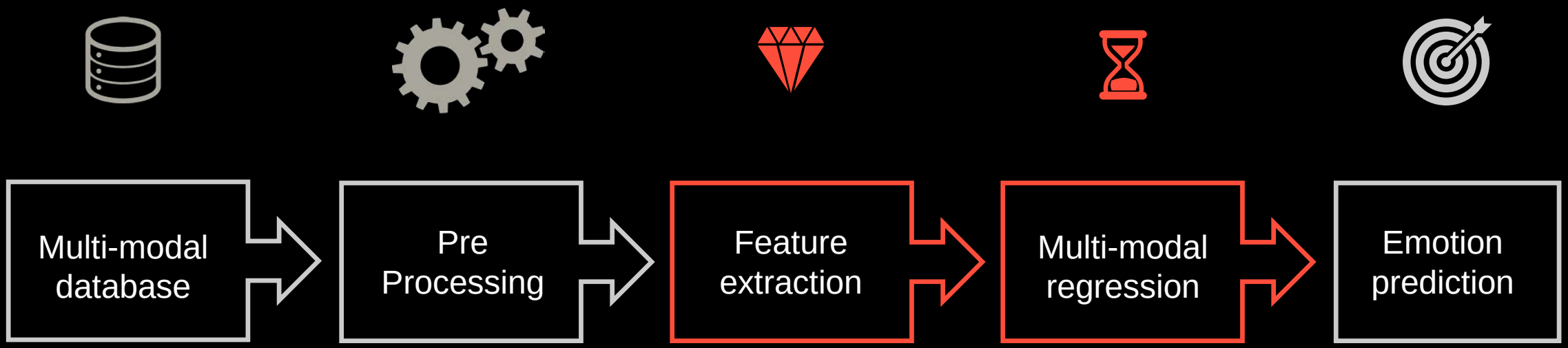
Features • Multi-modal benchmark • Qualitative study • Ablation study.

05 Conclusions

Conclusions • Future work.

General scheme

Emotion recognition systems can vary in their implementations but they can be summarized in a general scheme, as follows:



Multi-modal databases

MOSI

93 Youtube videos, 89 speakers.

Utterance annotation with **positive** and **negative** classification.

SEMAINE

24 videos role-played by humans.

Annotations every 0.2 seconds. Labels with valence and arousal.

IEMOCAP

10 actors split into pairs, 151 videos.

Annotations for anger, happiness, sadness, excitement, frustration, fear, surprise, neutral and other, at utterance level.

Feature Extraction

Context-free vectors (one-hot encoding) or context-based models (word embedding).

Segments, e.g. 30-200 milliseconds, are used to extract Low-level descriptors.

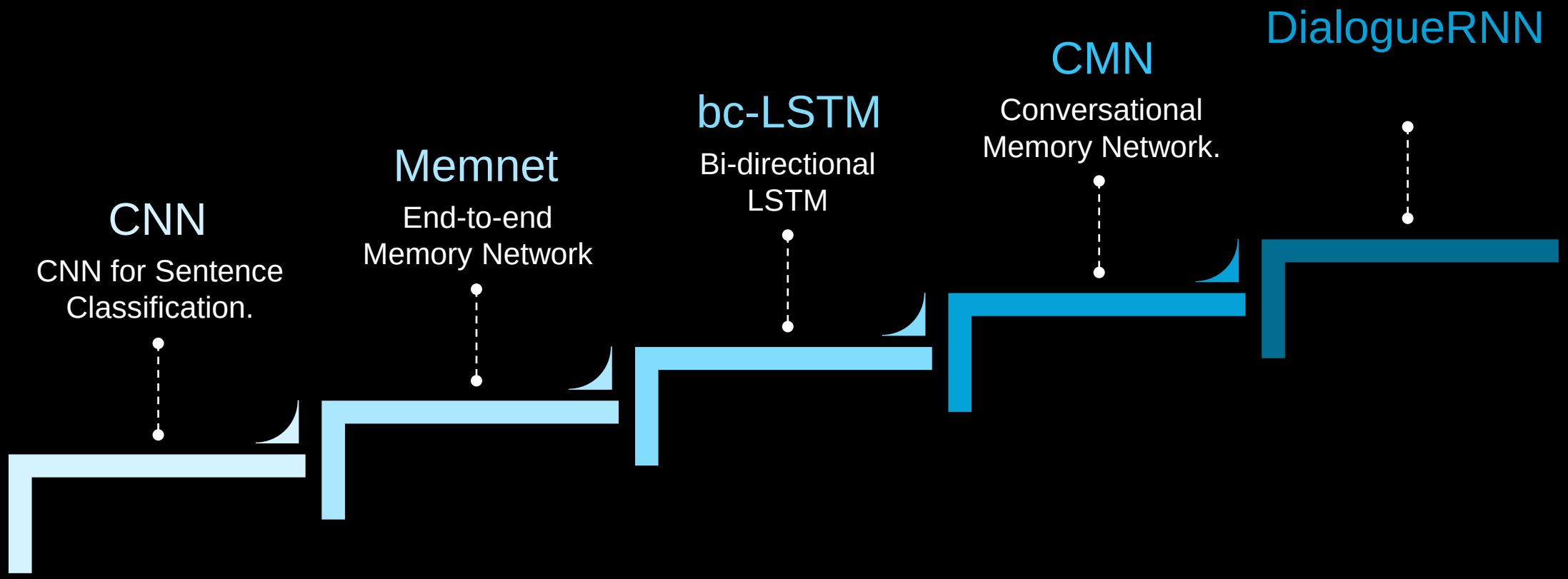
Deep features from audio could also be considered.

Videos are segmented into clips, e.g. 8-16 video frames, to obtain spatio-temporal features.

For each input modality, there is a specific neural network to be trained with the emotion labels. After training, the penultimate dense layer activations of these neural networks are used as a feature extractor.

Multi-modal methods

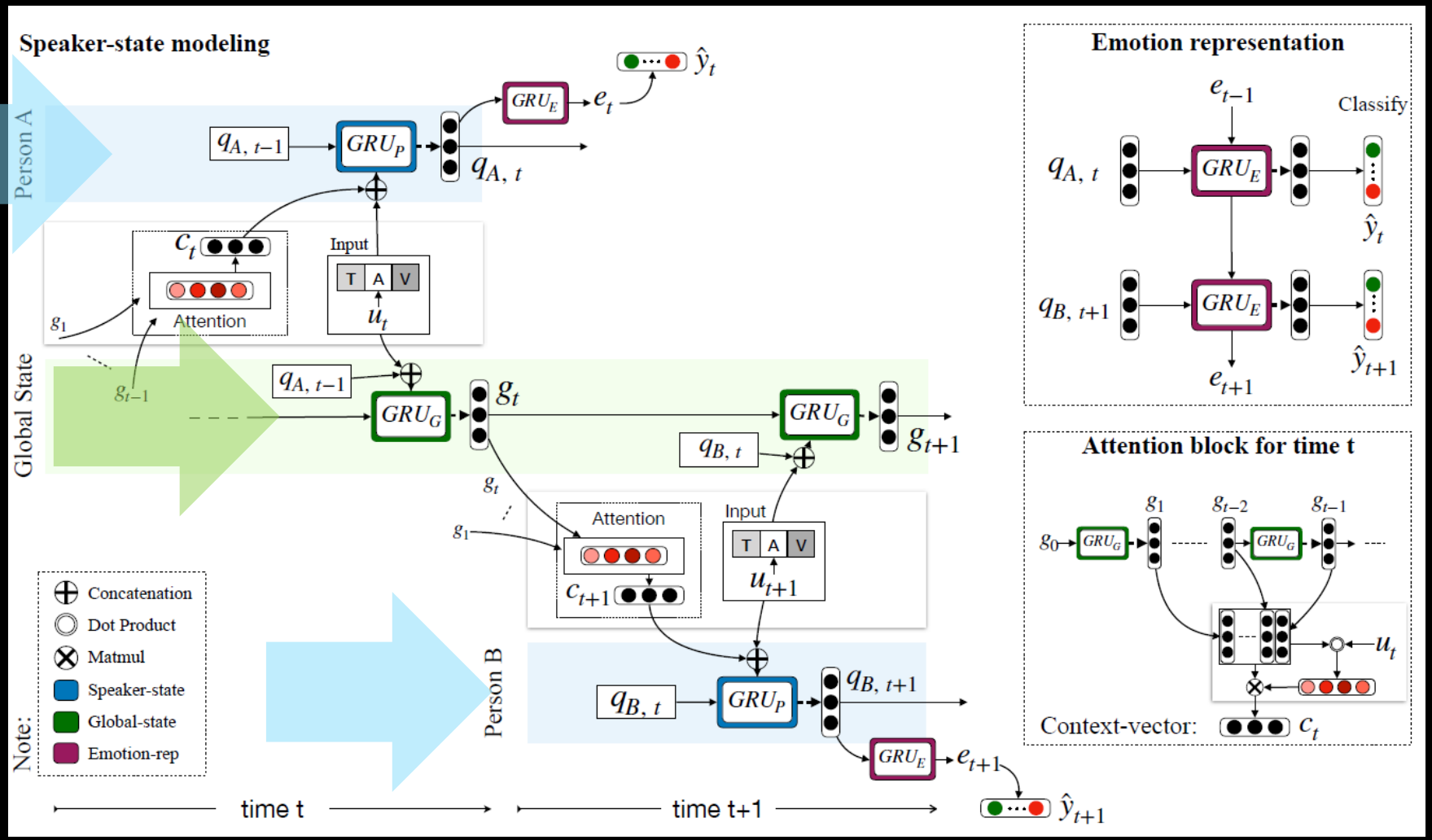
Extracted features for each modality are fed as inputs into multi-modal frameworks to add temporal information from previous utterances.



DialogueRNN

Three main components based on GRUs cells to describe temporal context and inter-speaker influence [1] :

1. Global state
2. Party state.
3. Emotion state.



[1] N. Majumder, S.Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. CoRR, abs/1811.00405, 2018. URL http://arxiv.org/abs/1811.00405.

Agenda

01 Introduction

Introduction • Objectives.

02 State of the art methods

Multi-modal Databases • Feature Extraction • Multi-modal methods

03 Proposal

Independent Feature Extraction • Joint Optimization

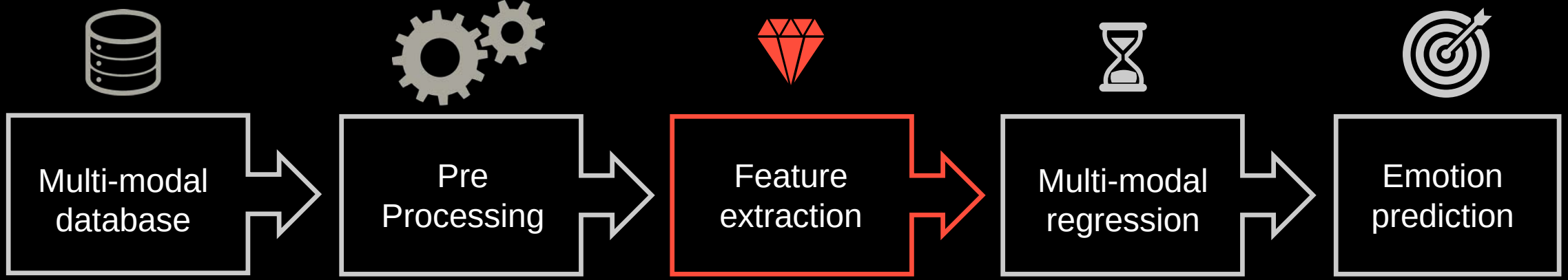
04 Experiments

Features • Multi-modal benchmark • Qualitative study • Ablation study.

05 Conclusions

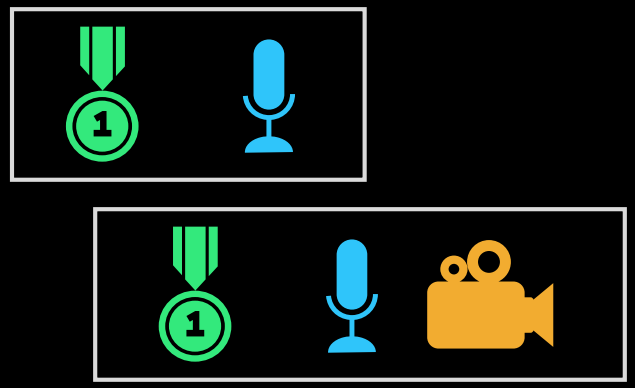
Conclusions • Future work.

Proposal

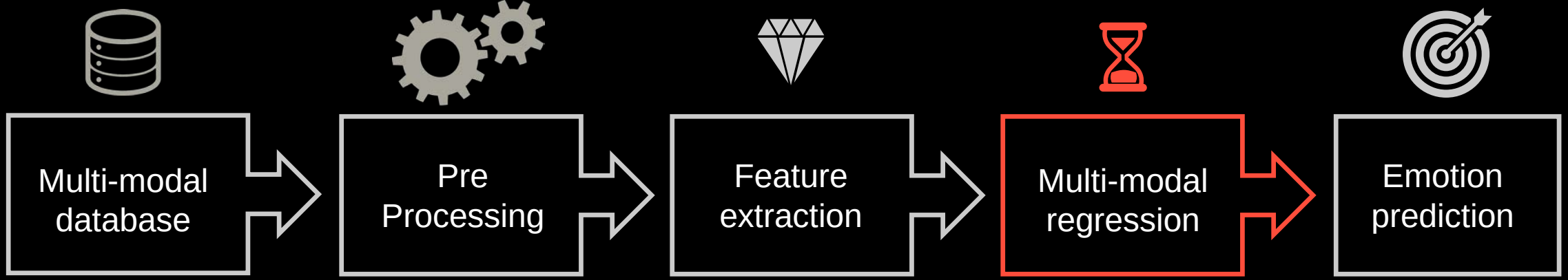


Independent feature extraction

Joint optimization of features



Proposal



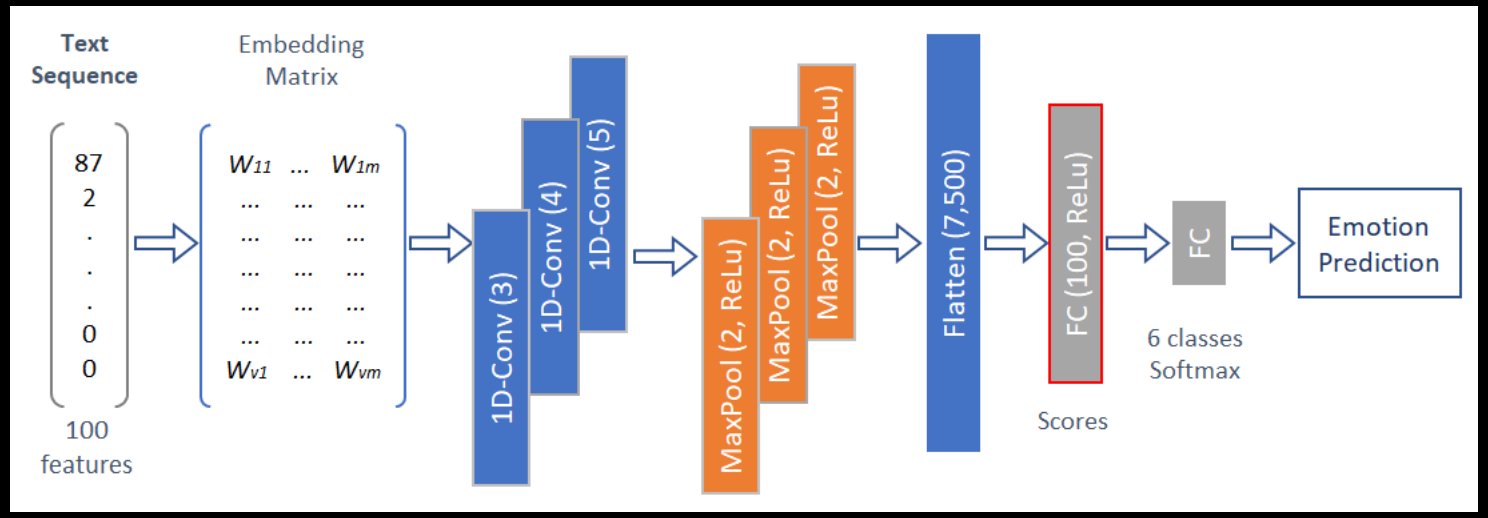
Baseline

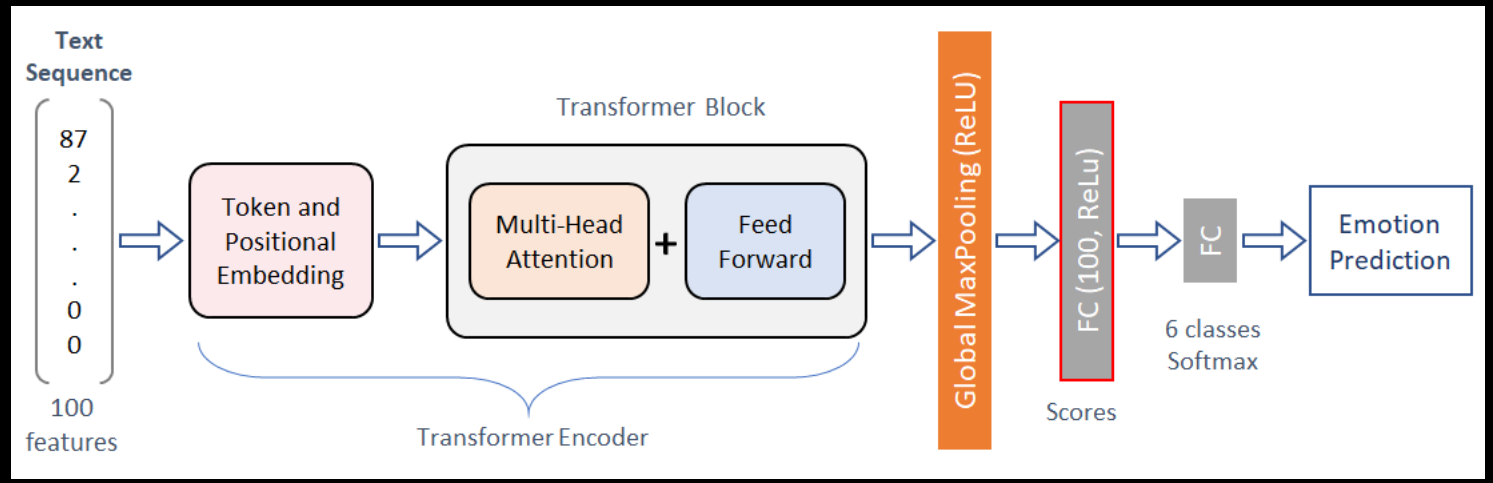
Transformers

Joint Optimization

Independent feature extraction - Text

Text based on CNN: includes pre-trained word embedding, 1D-CNN with filter sizes of 3, 4 and 5, max-pooling, flatten layer, dense layer with 100 neurons for “scores” and FC layers to output emotions.

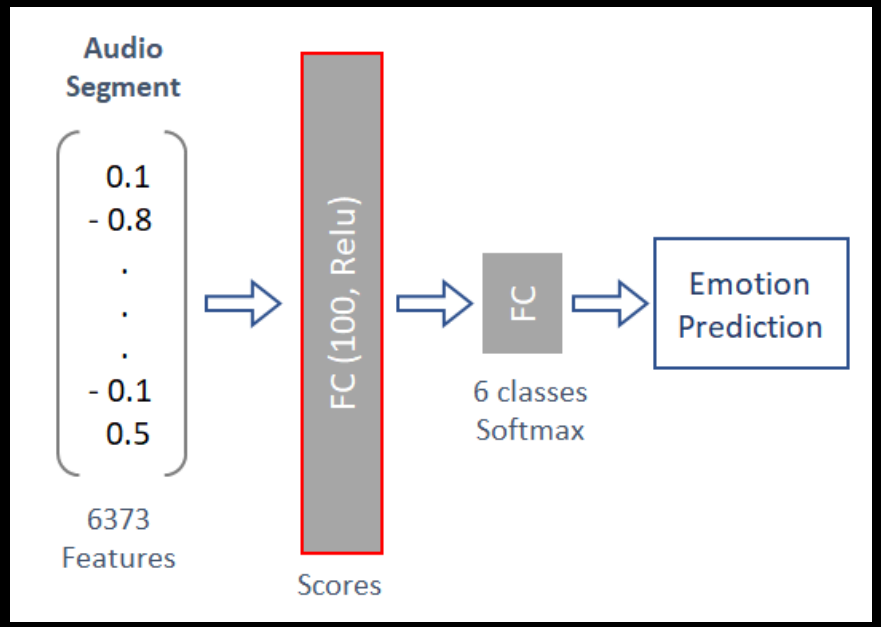




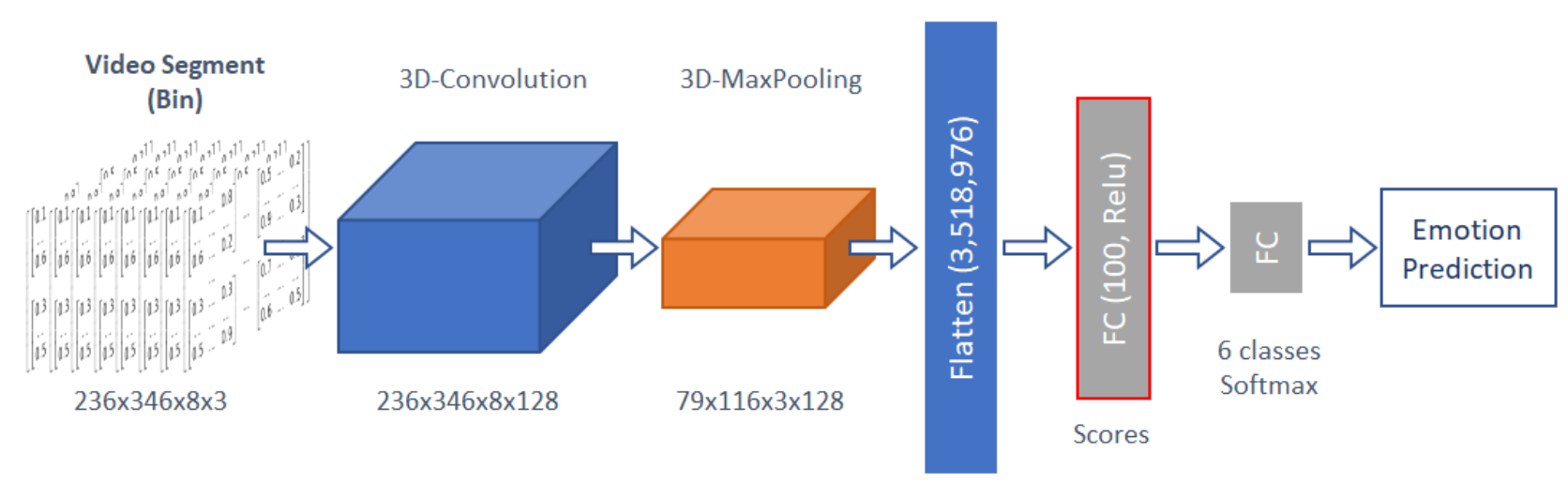
Text based on Transformer: Contains a token and positional embedding, transformer block with attention, a global average pooling, dense layer for “scores” with 100 neurons, and FC layer for emotion predictions.

Independent feature extraction - Audio

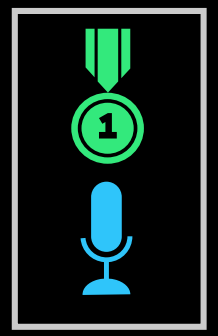
Input segments of Low-level descriptors with 6373 features extracted using OpenSMILE toolkit. The descriptors are fed through a dense layer with 100 neurons that serve as feature extractor, followed by a FC layer for emotion prediction.



Independent feature extraction - Video

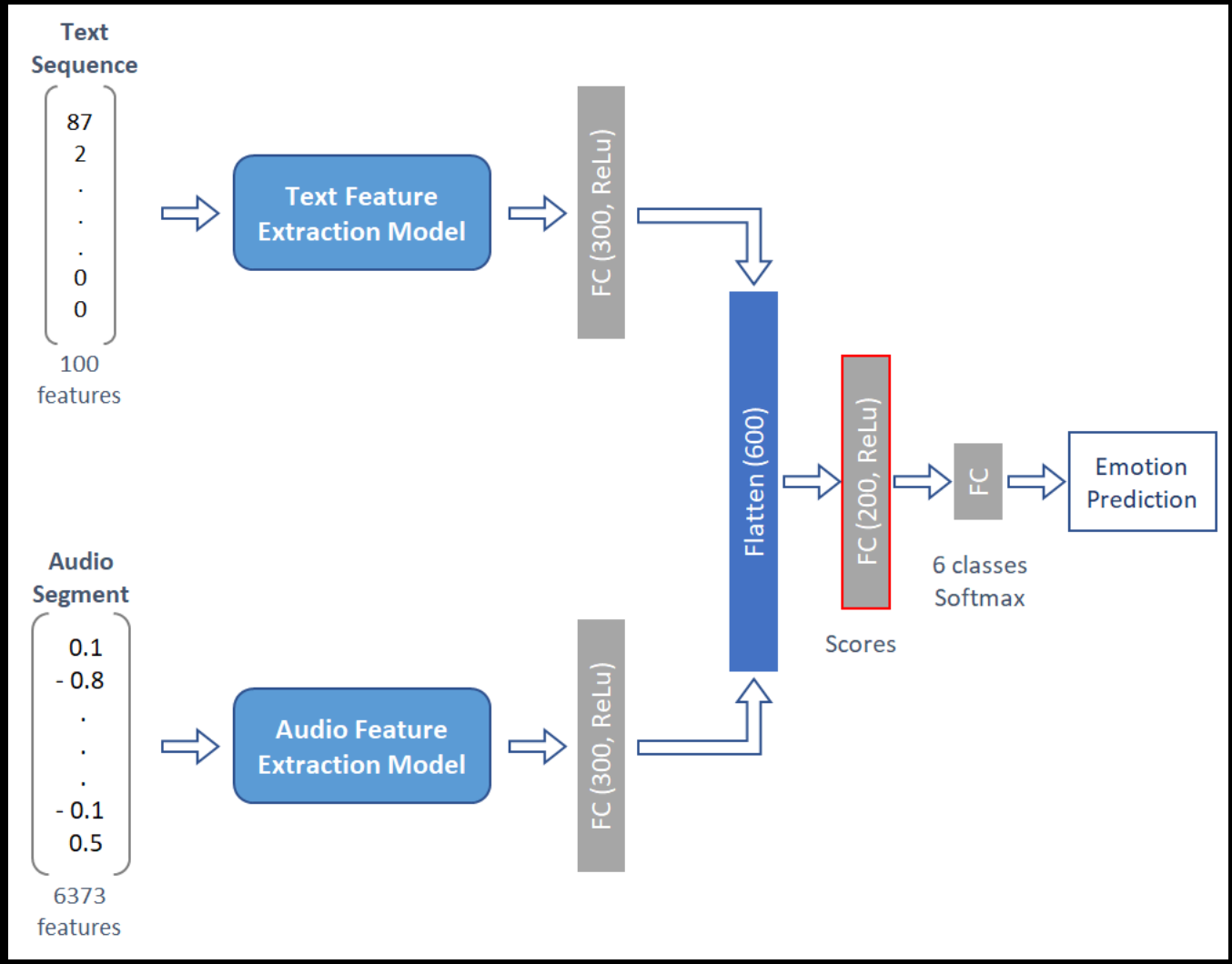


Input segments with 4 dimensions are linked to 3D-Convolution with 128 feature maps with size of 5, followed by 3D max-pooling with size of 3, a flatten layer and the 100 neurons dense layers for “scores”, to end with the FC layer for emotion prediction.

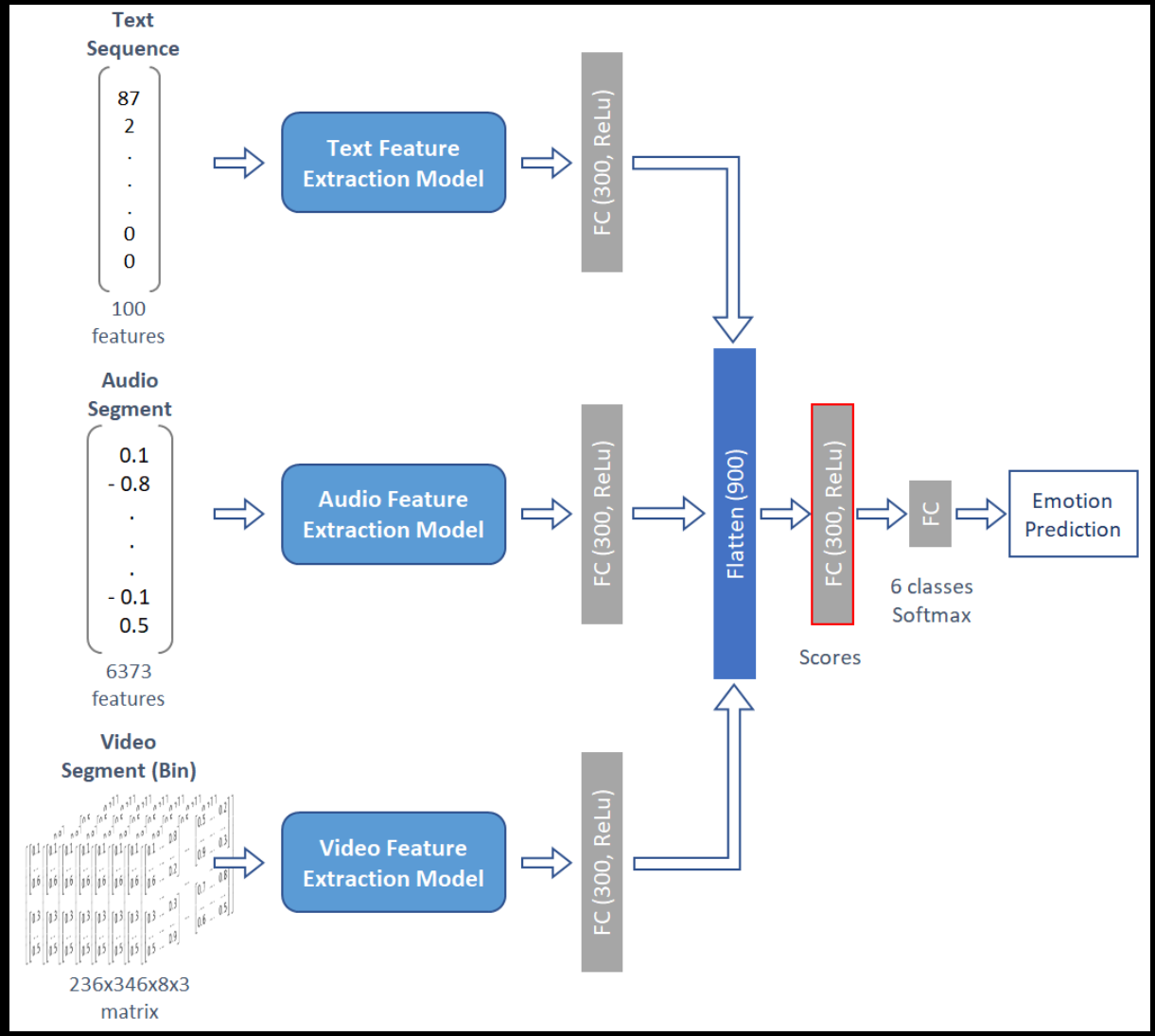


Joint optimization of features

Text and Audio: at the end of the text and audio feature extraction models are linked to a FC with 200 neurons. Both branches are joined with a flatten layer, followed by a dense layer with 100 neurons for “scores” and the FC layer for predictions.



Joint optimization of features



Text, Audio and Video: similar to the previous case with the addition of a new branch for video input modality. Also, the penultimate dense layer contains 300 neurons, 100 per each modality.

Agenda

01 Introduction

Introduction • Objectives.

02 State of the art methods

Multi-modal Databases • Feature Extraction • Multi-modal methods

03 Proposal

Independent Feature Extraction • Joint Optimization

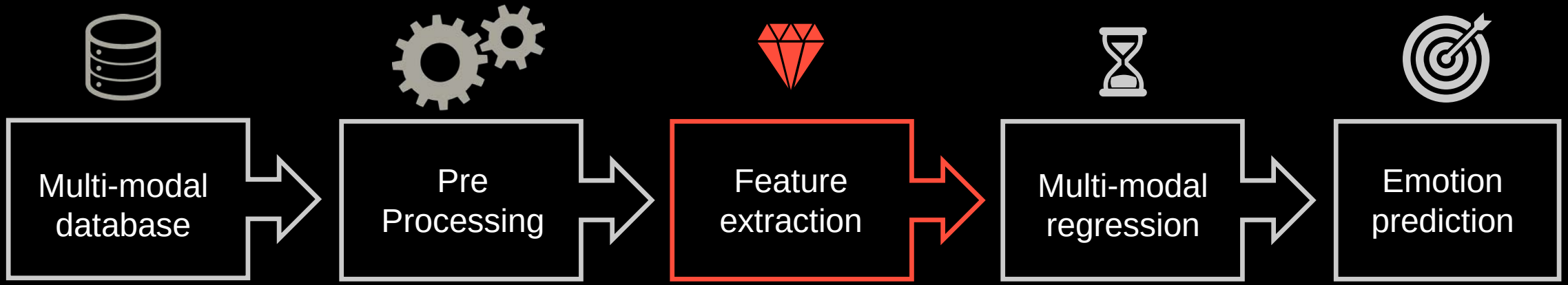
04 Experiments

Features • Multi-modal benchmark • Qualitative study • Ablation study.

05 Conclusions

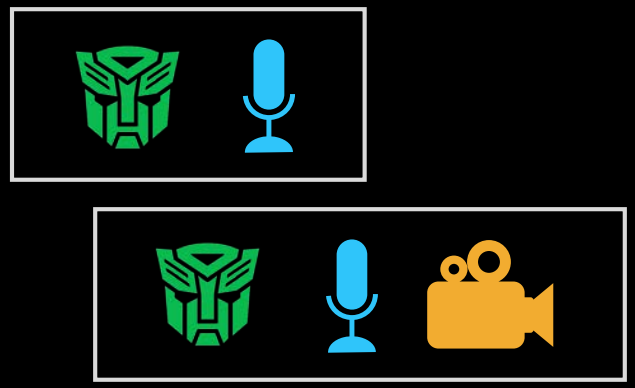
Conclusions • Future work.

Feature extraction



Independent feature extraction

Joint optimization of features



Independent feature extraction performance

Modalities	Output Dimensions	Number of parameters	Average (w)	
			Accuracy	F1-score
Text based on CNN	100	1 895 056	42.11	41.16
Text based on Transformers	100	1 407 738	**50.19**	**49.79**
Audio	100	638 206	24.93	9.97
Video	100	351 946 434	18.61	18.19

Best independent modality performance is obtained by text. In particular, text based on transformers enhance the baseline method using CNN.

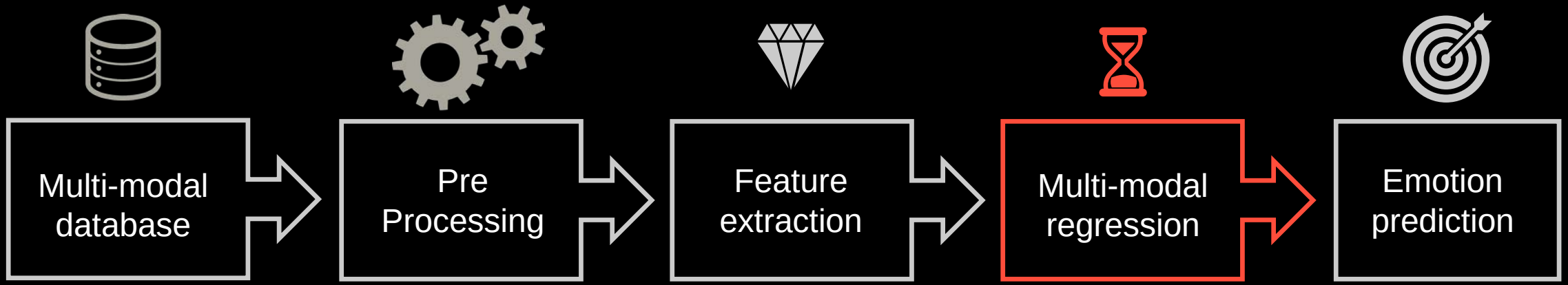
Video has the lowest average weighted accuracy and the largest architecture explained by the 3D-CNN layers.

Joint optimization performance

Modalities	Output Dimensions	Number of parameters	Average (w)	
			Accuracy	F1-score
Text$_{TE}$ + Audio	200	3 438 530	**49.73**	**49.30**
Text$_{TE}$ + Audio + Video	300	531 433 302	39.41	37.70

The additional branch of video increase the complexity of the model but it did not increase the performance. It is observed a compromise between the complexity and performance.

Multi-modal methods



Baseline

Transformers

Joint Optimization

Benchmark Multi-modal methods

Benchmark

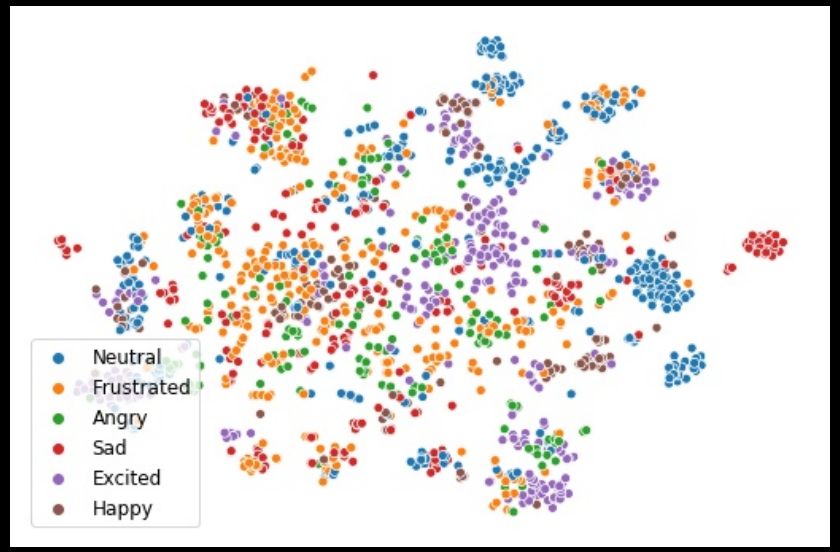
Methods	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average (w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	27.77	29.86	57.14	53.83	34.33	40.14	61.17	52.44	46.15	50.09	62.99	55.75	48.92	48.18
Memnet	25.72	33.53	55.53	61.77	58.12	52.84	59.32	55.39	51.50	58.30	67.20	59.00	55.72	55.10
bc-LSTM	29.17	34.43	57.14	60.87	54.17	51.81	57.06	56.73	51.17	57.95	67.19	58.92	55.21	54.95
bc-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19
CMN	25.00	30.38	55.92	62.41	52.86	52.39	61.76	59.83	55.52	60.25	**71.13**	60.69	56.56	56.13
DialogueRNN	31.25	33.83	66.12	69.83	**63.02**	**57.76**	61.76	62.50	61.54	64.45	59.58	59.46	59.33	59.89
DialogueRNN l	**35.42**	35.54	65.71	69.85	55.73	55.30	62.94	61.85	59.20	62.21	63.52	59.38	58.66	58.76
BiDialogueRNN	32.64	36.15	71.02	74.04	60.47	56.16	62.94	63.88	56.52	62.02	65.62	**61.73**	60.32	60.28
DialogueRNN+Att	28.47	**36.61**	65.31	72.40	62.50	57.21	**67.65**	**65.71**	70.90	68.61	61.68	60.80	61.80	61.51
BiDialogueRNN+Att	25.69	33.18	**75.10**	**78.80**	58.59	59.21	64.71	65.28	**80.27**	**71.86**	61.15	58.91	**63.40**	**62.75**

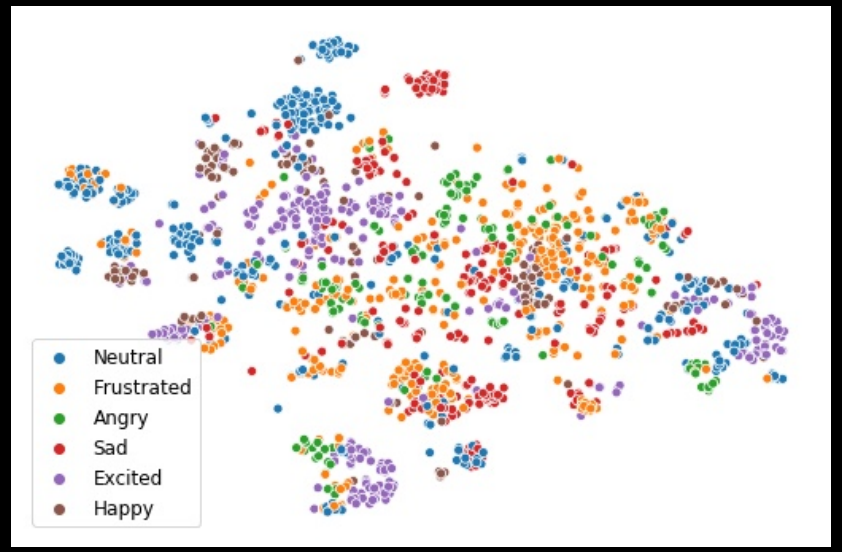
Proposed

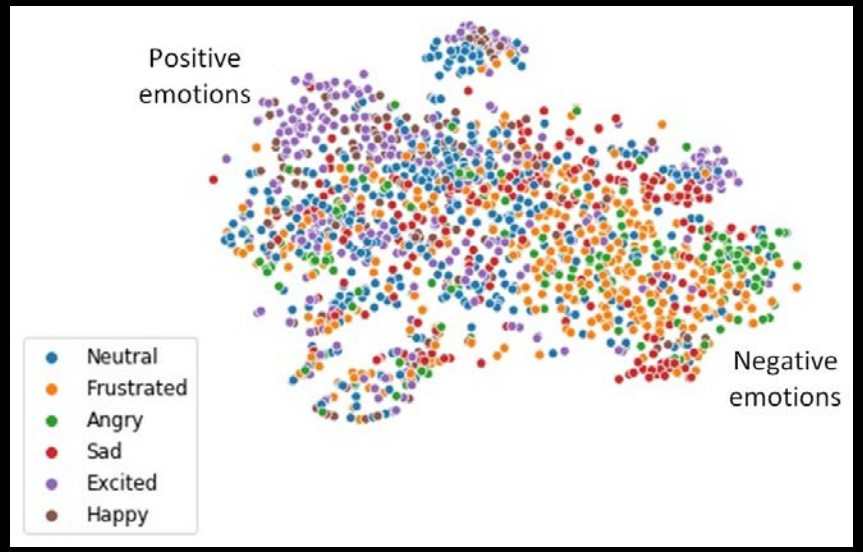
Methods	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average (w)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Baseline	32.87	29.84	15.10	21.70	55.73	37.06	**74.12**	44.37	11.71	18.09	7.35	11.72	30.02	25.42
Transformers	29.37	35.00	**18.78**	**28.40**	77.60	48.06	55.88	**46.12**	27.42	41.52	28.35	34.12	41.37	**39.26**
Joint Optimization	**37.76**	**43.72**	7.76	14.18	**85.68**	**48.28**	0.59	1.17	**34.45**	**48.13**	**49.87**	**49.54**	**42.91**	38.06

Qualitative study

The performance benchmark is complemented with a qualitative analysis using t-SNE algorithm for dimensionality reduction of each multi-modal proposed method.







Acc./F1-score: 30.02 / 25.42

41.37 / **39.26**

42.91 / 38.06

Baseline y Transformers:

- Multiple and small clusters for each emotion label.
- Mixed emotions in the middle.

Joint optimization:

- Emotions have few number of clusters.
- Similar emotions are close to each other.

Ablation study

Methods	Modalities	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average (w)	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Baseline	TextCNN	60.84	46.65	70.20	73.35	56.25	52.11	53.53	54.98	52.51	64.08	51.44	52.13	56.66	57.35
	Audio	34.27	27.15	48.16	29.72	1.56	2.95	6.47	8.87	0.00	0.00	47.51	31.89	22.50	16.00
	Video	0.00	0.00	42.04	37.66	1.04	1.85	31.18	16.09	45.48	27.45	0.00	0.00	18.25	12.87
	TextCNN + Audio	33.57	34.41	**86.53**	78.23	61.72	**58.02**	49.41	55.63	65.55	65.66	48.03	51.77	59.19	58.68
	TextCNN + Video	29.37	38.89	25.31	32.80	33.59	39.45	**85.88**	28.27	25.75	26.46	0.00	0.00	28.11	25.56
	Audio + Video	0.00	0.00	0.00	0.00	0.00	0.00	68.24	26.76	41.14	29.25	27.56	27.49	21.21	14.65
	TextCNN + Audio + Video	32.87	29.84	15.10	21.70	55.73	37.06	74.12	44.37	11.71	18.09	7.35	11.72	30.02	25.42
Transformer	TextTE	51.05	46.50	84.90	82.87	57.03	57.33	64.12	**56.48**	66.56	68.62	48.56	53.01	61.22	61.21
	TextTE + Audio	62.94	**56.43**	**86.53**	**83.79**	50.78	56.20	53.53	54.49	**75.25**	**75.00**	**60.37**	**58.15**	**64.30**	**64.13**
	TextTE + Video	30.07	33.08	51.84	58.12	64.84	45.90	47.65	47.09	24.08	37.31	38.06	39.62	44.20	43.68
	TextTE + Audio + Video	29.37	35.00	18.78	28.40	77.60	48.06	55.88	46.12	27.42	41.52	28.35	34.12	41.37	39.26
Joint Optimization	TextTE + Audio	**77.62**	52.24	85.71	81.08	55.73	57.37	57.65	53.70	51.51	64.57	48.03	51.33	59.80	60.02
	TextTE + Audio + Video	37.76	43.72	7.76	14.18	**85.68**	48.28	0.59	1.17	34.45	48.13	49.87	49.54	42.91	38.06

Text with transformer is able to reach 61.22 and 61.21 in average weighted accuracy and F1-score, respectively. Moreover, text with transformers + audio, outperforms DialogueRNN best multi-modal method by 0.90 and 1.38 percentage points in average weighted accuracy and F1-score, each.

Agenda

01 Introduction

Introduction • Objectives.

02 State of the art methods

Multi-modal Databases • Feature Extraction • Multi-modal methods

03 Proposal

Independent Feature Extraction • Joint Optimization

04 Experiments

Features • Multi-modal benchmark • Qualitative study • Ablation study.

05 Conclusions

Conclusions • Future work.

Conclusions

- Obj. 1-3: data, implementation and benchmark comparison.
- Obj. 4: alternative to text → enhance feature representation.
- Obj. 5: joint optimization of features → Latent factors in emotions.
- Obj. 6: modality importance evaluation → text transformers & audio outperforms state-of-the-art.

Future work

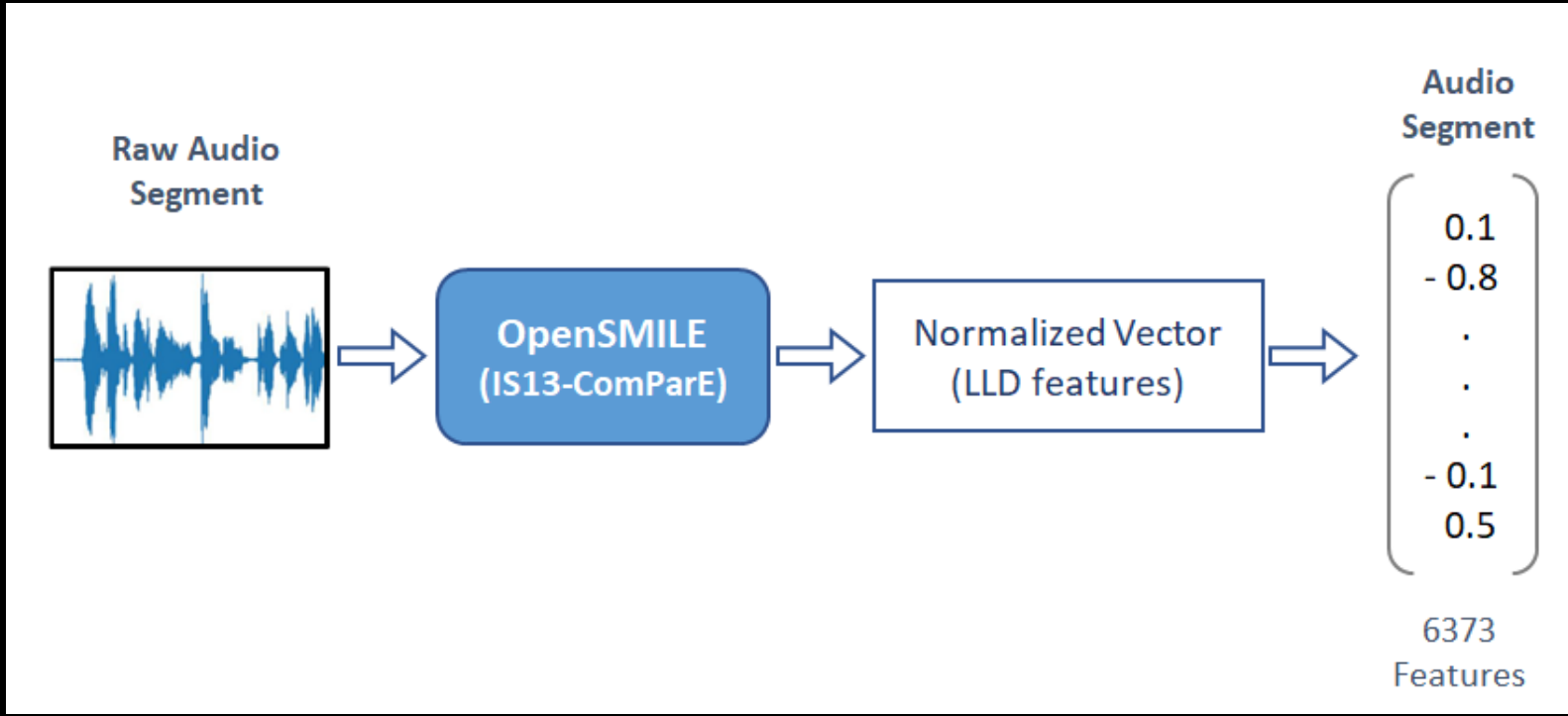
- Improve audio and video features → data augmentation / graph or pruning.
- In depth study of emotions relationships.

Deep Regression of Social Signals in Dyadic Scenarios

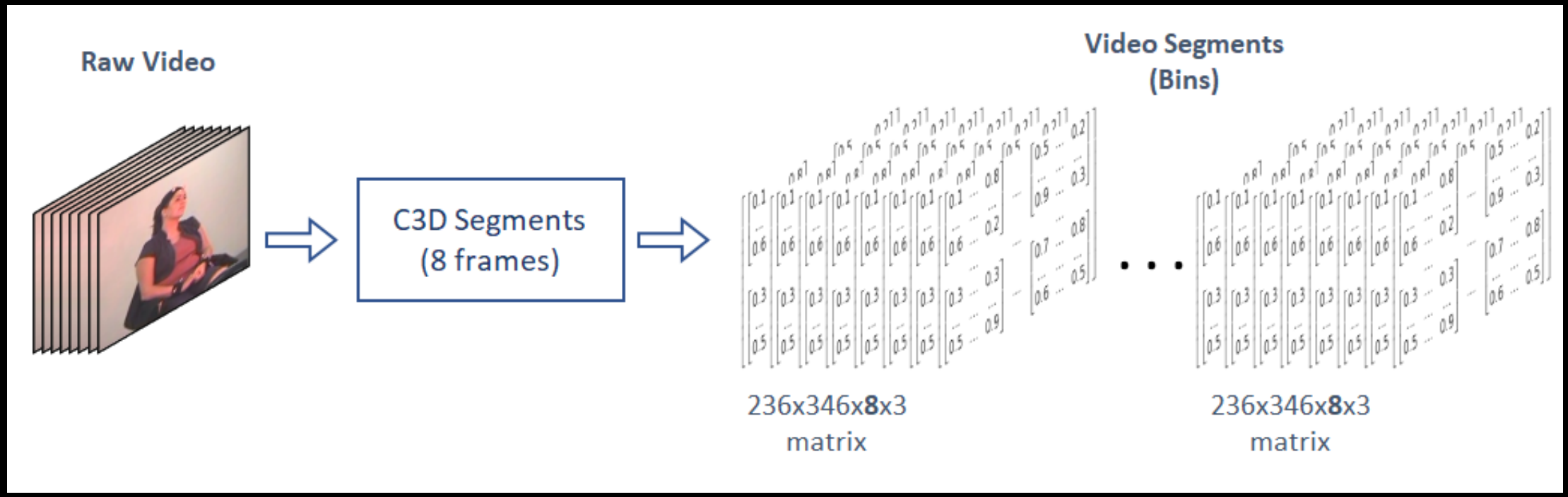
Author: Ítalo Vidal Lucero
Director: Dr. Sergio Escalera
Co-directors: Dr. Julio Jacques & Cristina Palmero

Annexes

Audio pre-processing



Video pre-processing



Transformer architecture