Tensor Fusion Network for Multimodal Sentiment Analysis

Sanghyuck Na
May, 5, 2021
Dongguk University
Artificial Intelligence Laboratory
shna@Dongguk.edu

0

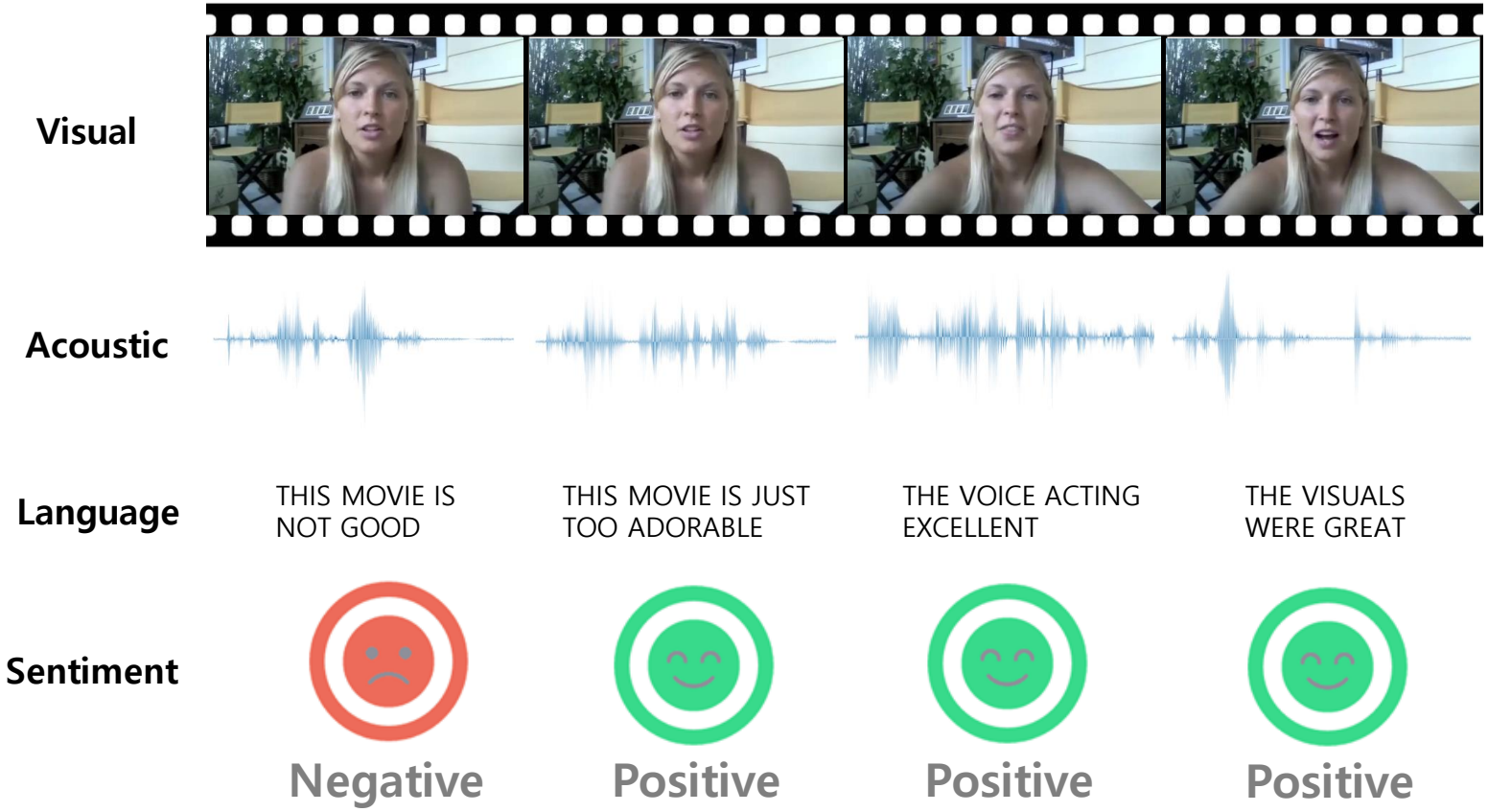
Contents

1 Multimodal Sentiment Analysis

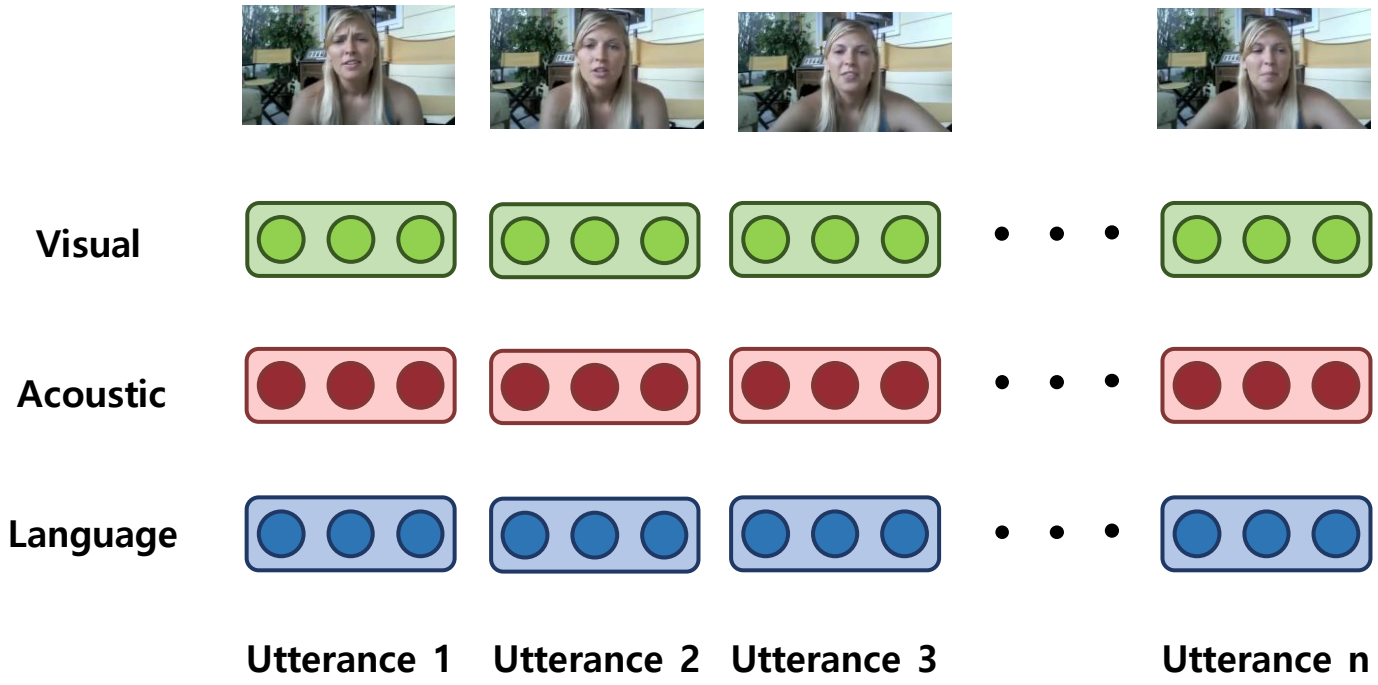
Loves the German bakeries in Sydney. Together with my imported honey it feels like home	Positive
@VivaLaLauren Mine is broken too! I miss my sidekick	Negative
Finished fixing my twitter... I had to unfollow and follow everyone again	Negative
@DinahLady I too, liked the movie! I want to buy the DVD when it comes out	Positive
@frugaldougal So sad to hear about @OscarTheCat	Negative
@Mofette brilliant! May the fourth be with you #starwarsday #starwars	Positive
Good Morning thespians a bright and sunny day in UK, Spring at last	Positive
@DowneyisDOWNEY Me neither! My laptop's new, has dvd burning/ripping software but I just can't copy the files somehow!	Negative

Text Sentiment Analysis

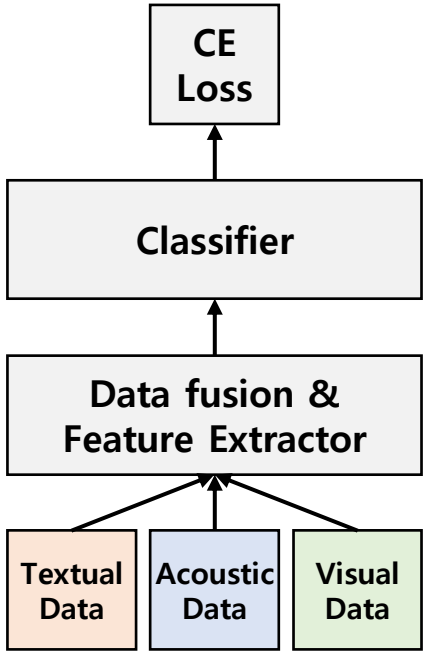
1 Multimodal Sentiment Analysis



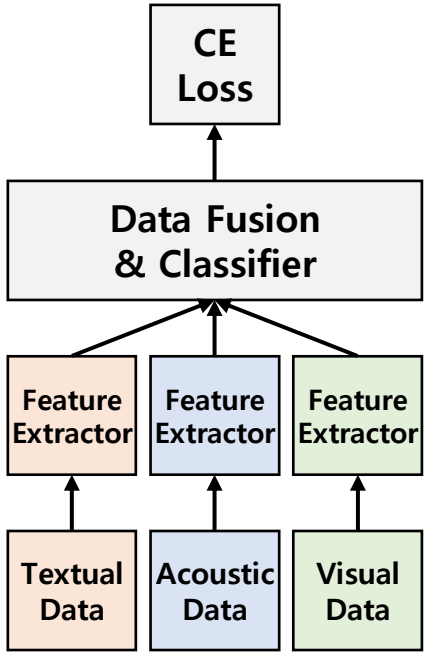
1 Multimodal Sentiment Analysis



1 Multimodal Sentiment Analysis



(a) Early Fusion based
Multimodal Sentiment Analysis



(b) Late Fusion based
Multimodal Sentiment Analysis

MMMU-BA

MMMU-BA(Multi-Modal Multi-Utterance Bi-modal Attention)

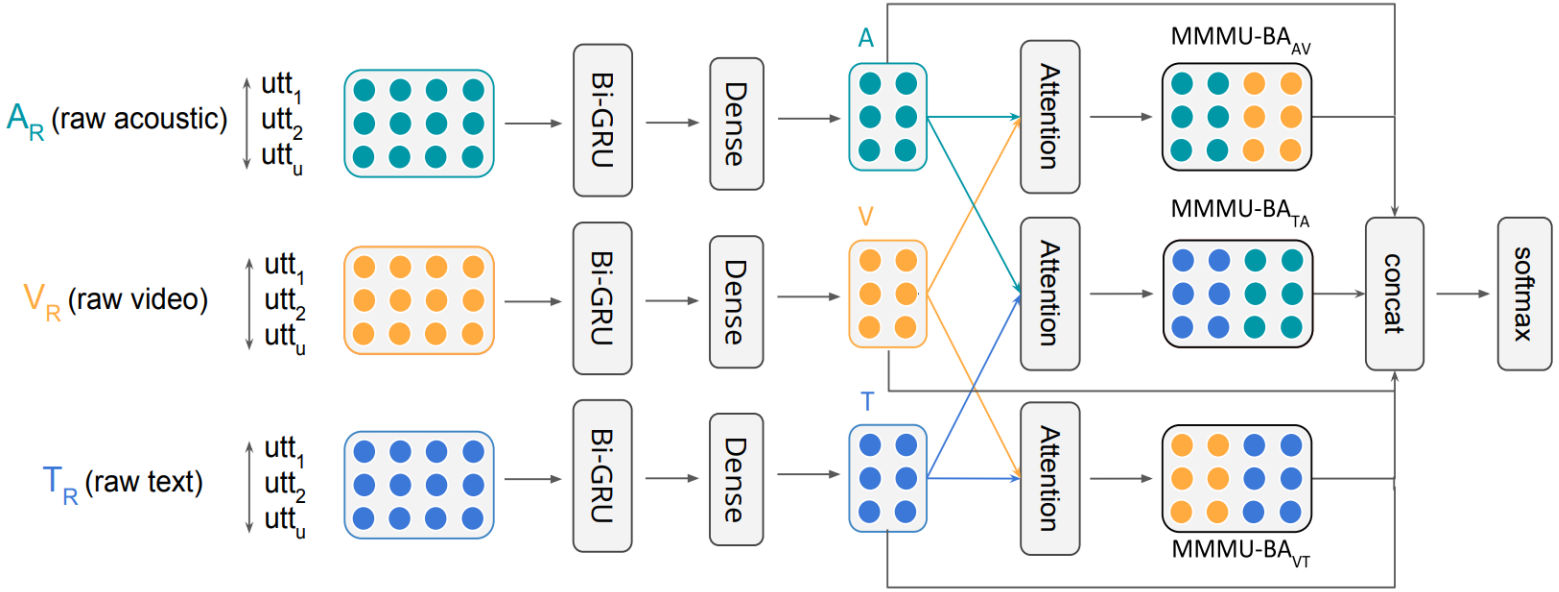


Figure 1: Overall architecture of the proposed MMMU-BA framework.

This is a model that obtains bimodal attention from between two or more modalities to perform the final classification

2 MMMU-BA

MU-SA(Multi-Utterance Self Attention)

The MU-SA model applies self-attention to the extracted features.

```
m = layers.dot([features, features], axes=[2, 2])
n = self.softmax(m)
o = layers.dot([n, features], axes=[2, 1])
a = layers.multiply([o, features])
```

MMUU-SA(Multi-Modal Uni-Utterance Self Attention)

This is a methodology that utilizes the characteristics of two or more modalities such as MMMU-BA, but uses the self-attention of each modality without considering the attention among modalities

```
attention_features = []
    for k in range(max_utt_len):
    # extract multi modal features for each utterance #

        m1 = feature_list[1][:, k:k + 1, :]
        m2 = feature_list[2][:, k:k + 1, :]
        m3 = feature_list[0][:, k:k + 1, :]

        utterance_features = layers.concatenate([m1, m2, m3], axis=1)
        attention_features.append(self.self_attention(utterance_features))

    merged_attention = layers.concatenate(attention_features, axis=1)
    merged_attention = tf.reshape(merged_attention,
                                  ((-1, max_utt_len, 3 * self.text_dense_units)))

    merged_features = layers.concatenate([
                        merged_attention, feature_list[1], feature_list[2], feature_list[0]])
```

Tensor Fusion Networks

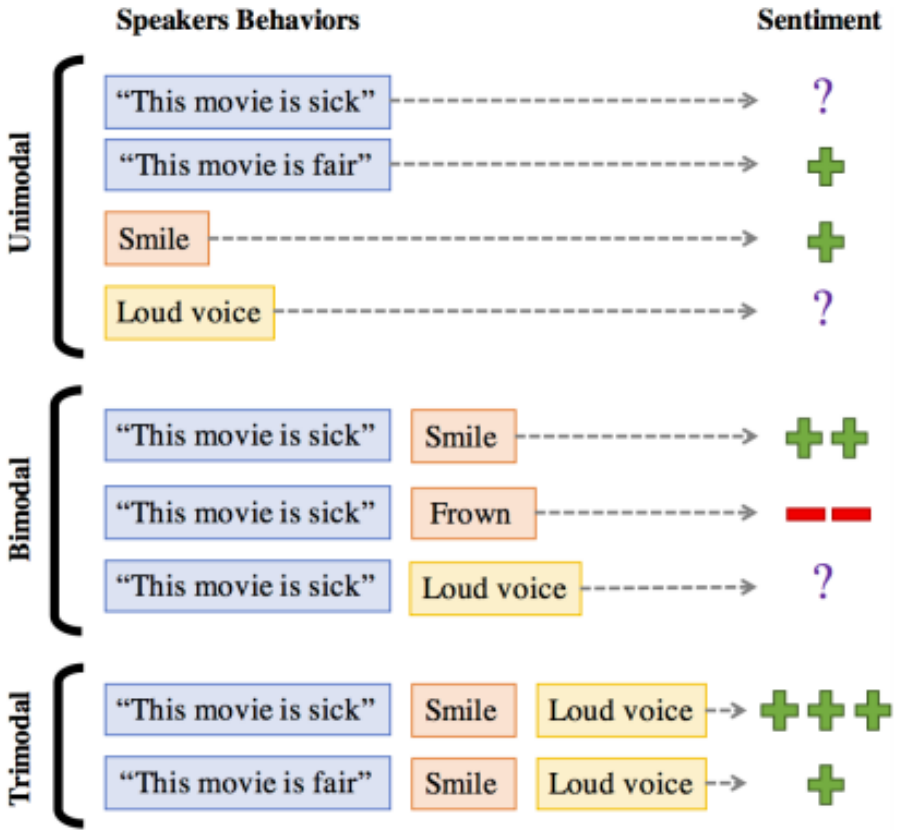
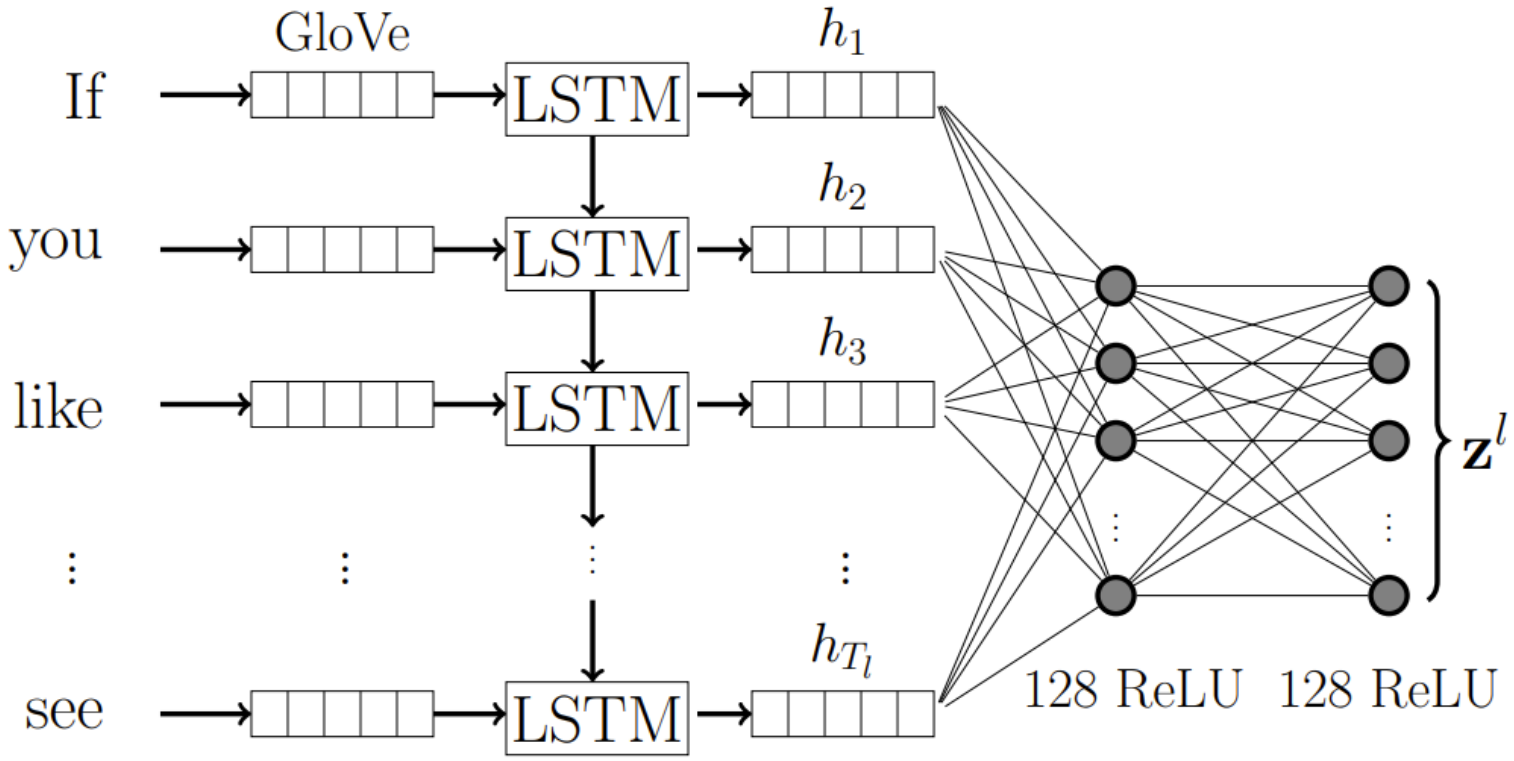


Figure 1: Unimodal, bimodal and trimodal interaction in multimodal sentiment analysis.

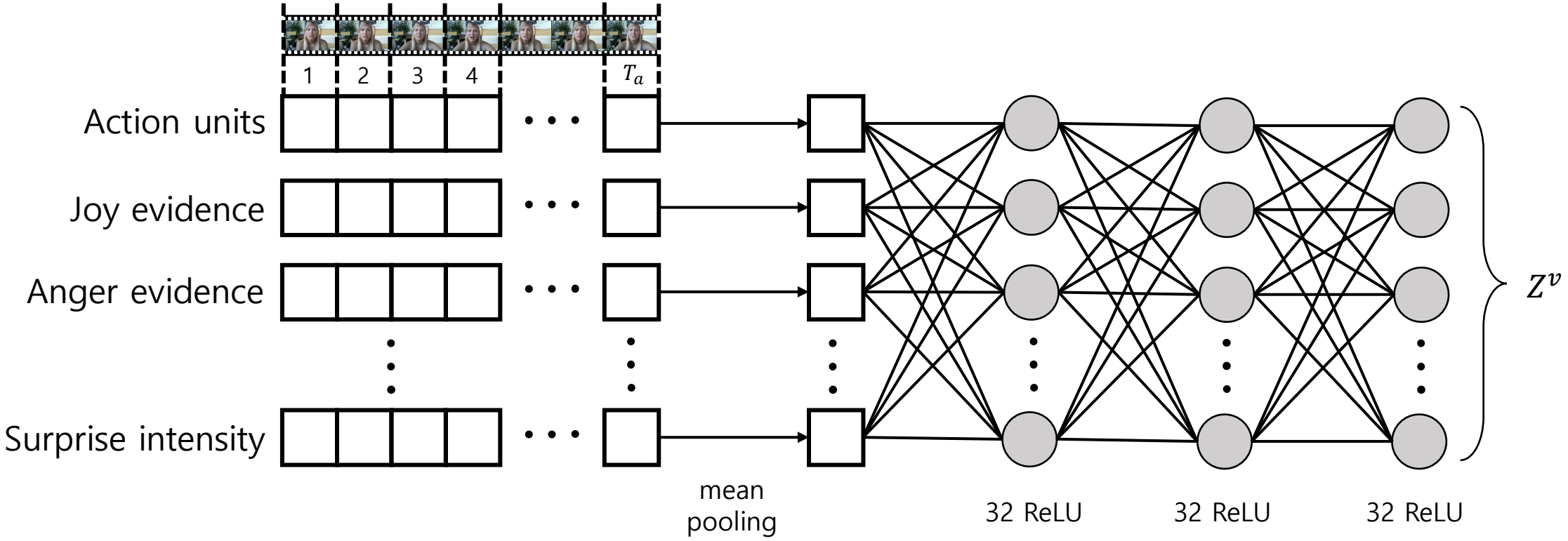
3 Tensor Fusion Networks

Spoken Language Embedding Subnetwork



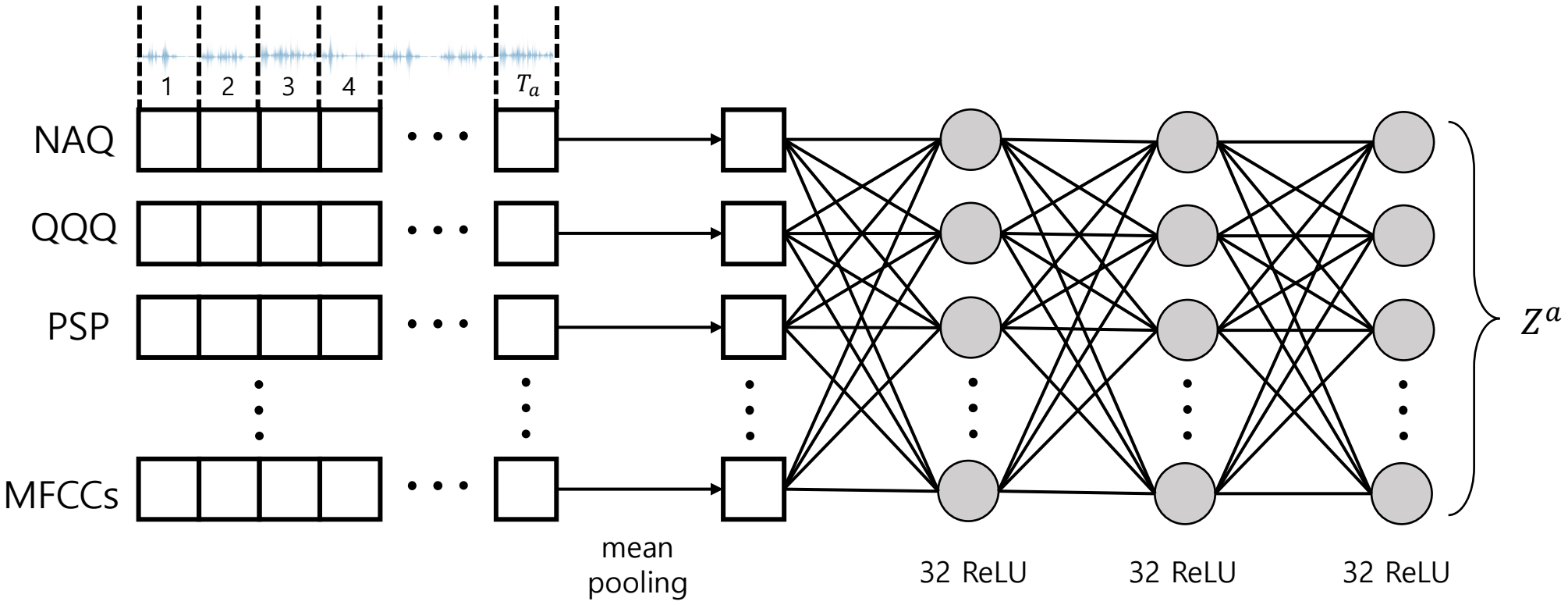
3 Tensor Fusion Networks

Visual Embedding Subnetwork using FACET framework

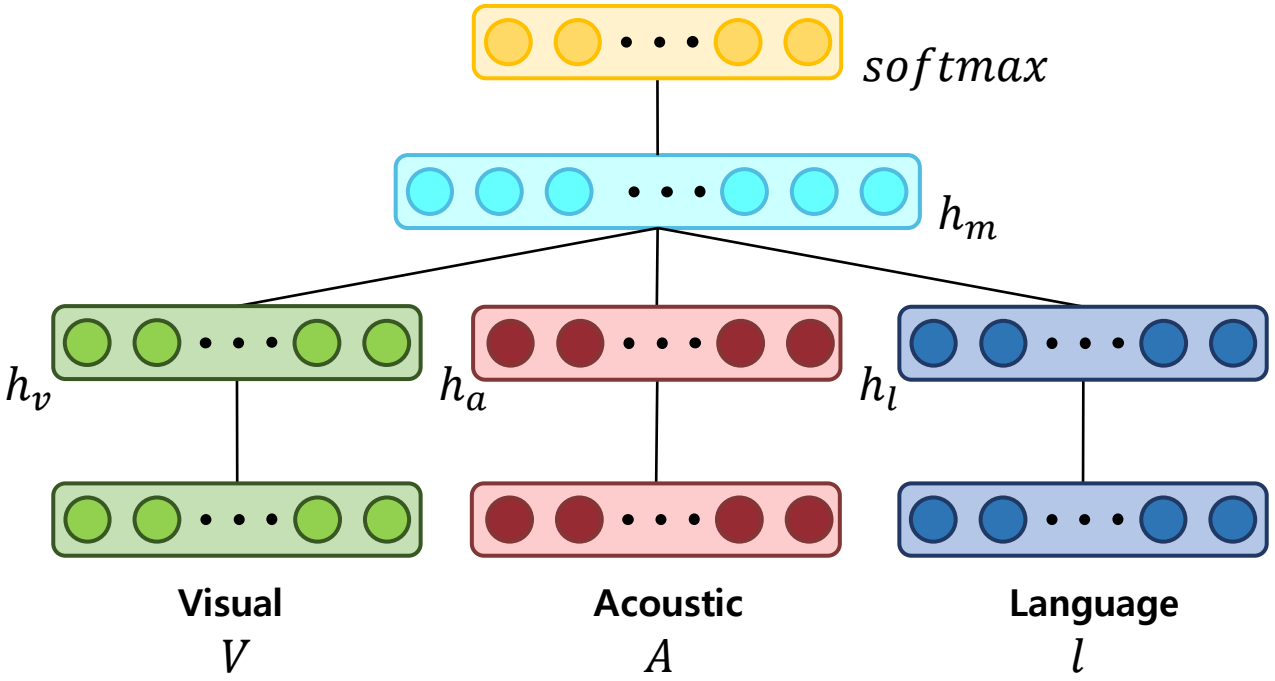


3 Tensor Fusion Networks

Acoustic Embedding Subnetwork using COVAREP framework



3 Tensor Fusion Networks

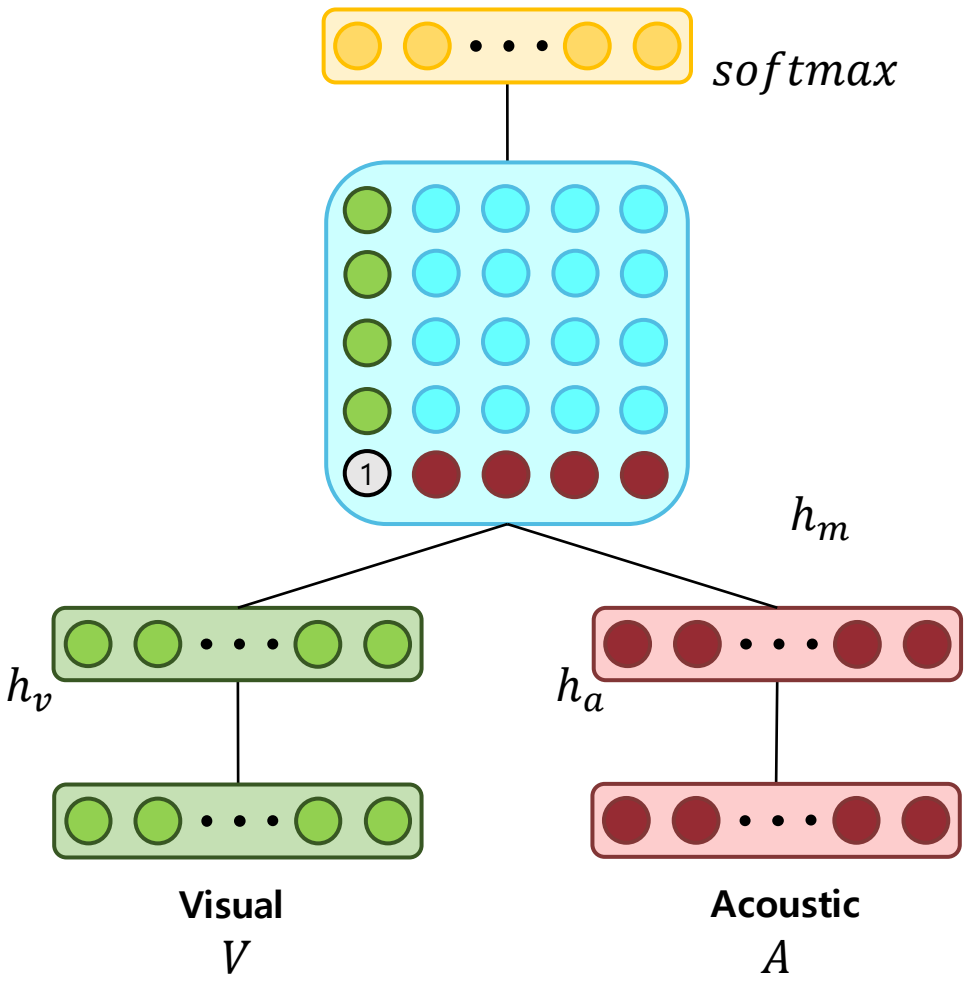


Joint Multimodal Representation :

Simply concatenates all three individual representation

$$h_m = f(W \cdot [h_v, h_a, h_l])$$

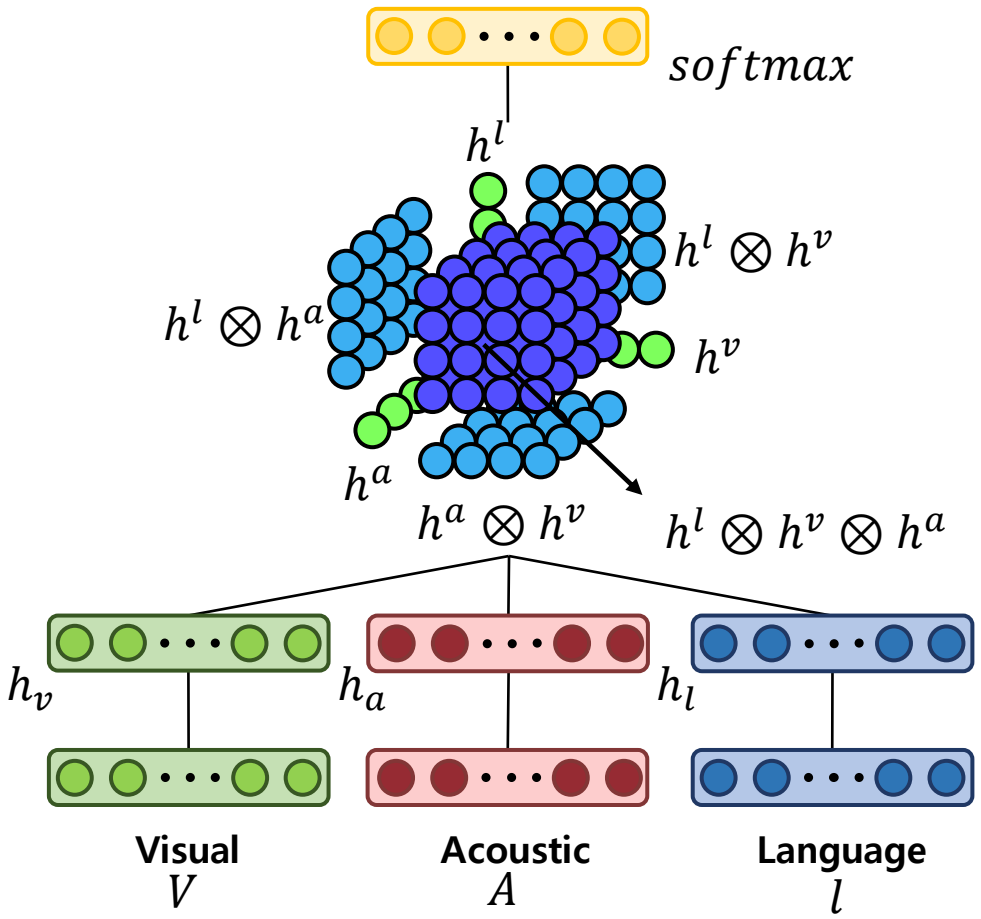
3 Tensor Fusion Networks



Multimodal Tensor Fusion Network for Bimodal

$$h_m = \begin{bmatrix} h_v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_a \\ 1 \end{bmatrix} = \begin{bmatrix} h_v & h_v \otimes h_a \\ 1 & h_a \end{bmatrix}$$

3 Tensor Fusion Networks



Multimodal Tensor Fusion Network for Trimodal

$$h_m = \begin{bmatrix} h_v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_a \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_l \\ 1 \end{bmatrix}$$

3 Tensor Fusion Networks

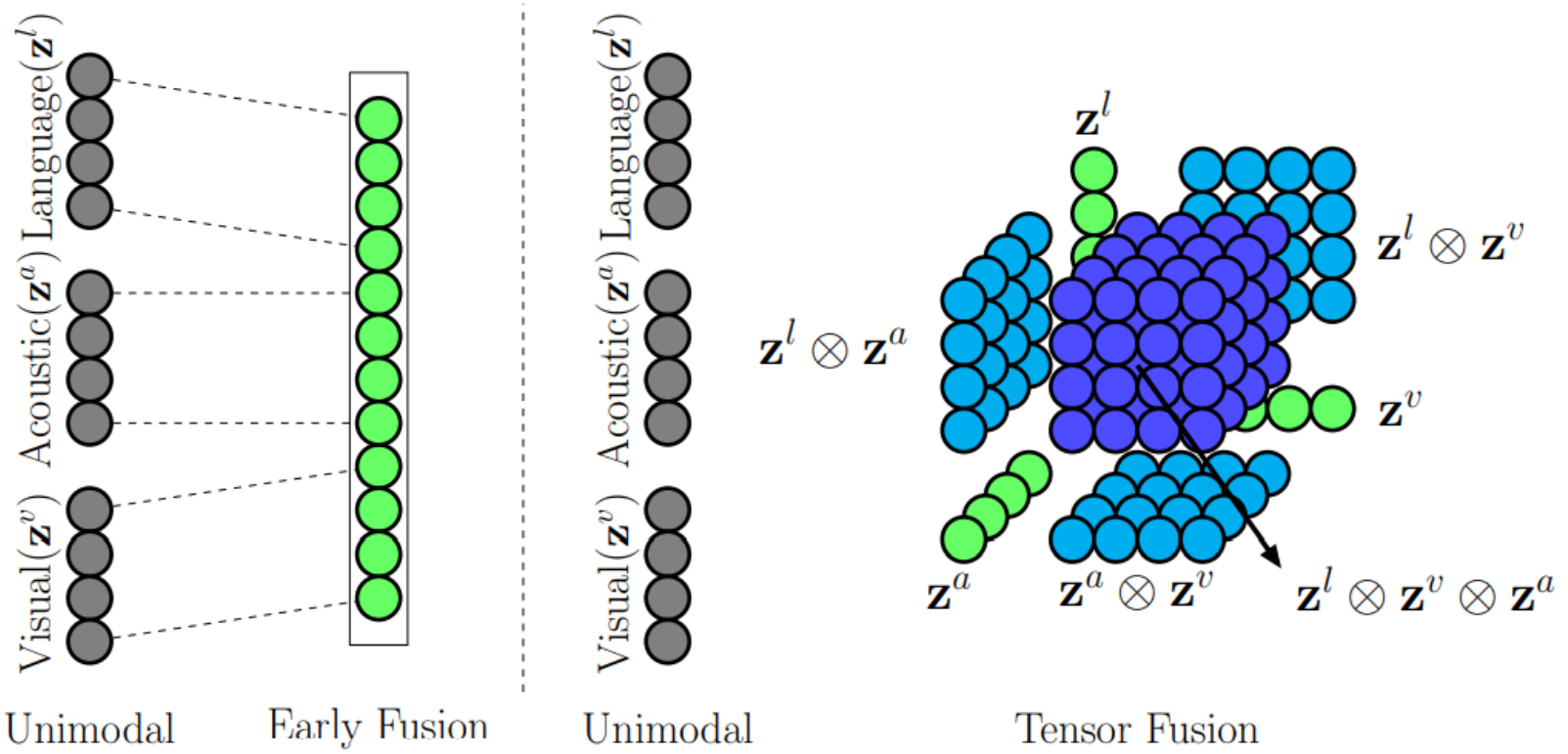


Figure 4: Left: Commonly used early fusion (multimodal concatenation). Right: Our proposed tensor fusion with three types of subtensors: unimodal, bimodal and trimodal.

Result

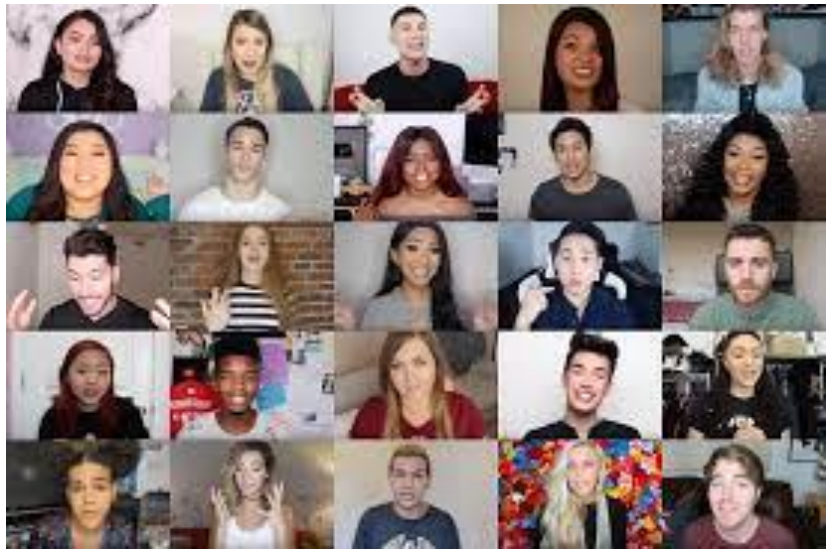
CMU-MOSI Dataset

3 prediction tasks

1. Binary classification
   - Positive or Negative

2. 5-class classification
   - Positive, Weakly Positive, Neutral, Weakly Negative, Negative

3. Regression
   - From -3 to +3



4 Result

Multimodal Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
Random	50.2	48.7	23.9	1.88	-
C-MKL	73.1	75.2	35.3	-	-
SAL-CNN	73.0	-	-	-	-
SVM-MD	71.6	72.3	32.0	1.10	0.53
RF	71.4	72.1	31.9	1.11	0.51
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
Human	85.7	87.5	53.9	0.71	0.82
Δ^{SOTA}	↑ 4.0	↑ 2.7	↑ 6.7	↓ 0.23	↑ 0.17

Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
$\text{TFN}_{language}$	74.8	75.6	38.5	0.99	0.61
TFN_{visual}	66.8	70.4	30.4	1.13	0.48
$\text{TFN}_{acoustic}$	65.1	67.3	27.5	1.23	0.36
$\text{TFN}_{bimodal}$	75.2	76.0	39.6	0.92	0.65
$\text{TFN}_{trimodal}$	74.5	75.0	38.9	0.93	0.65
$\text{TFN}_{notrimodal}$	75.3	76.2	39.7	0.919	0.66
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
TFN_{early}	75.2	76.2	39.0	0.96	0.63

4 Result

Language Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
RNTN	- (73.7)	- (73.4)	- (35.2)	- (0.99)	- (0.59)
DAN	73.4 (68.8)	73.8 (68.4)	**39.2** (36.7)	- -	- -
D-CNN	65.5 (62.1)	66.9 (56.4)	32.0 (32.4)	- -	- -
CMKL-L	71.2	72.4	34.5	-	-
SAL-CNN-L	73.5	-	-	-	-
SVM-MD-L	70.6	71.2	33.1	1.18	0.46
$\text{TFN}_{language}$	**74.8**	**75.6**	38.5	**0.98**	**0.62**
$\Delta^{SOTA}_{language}$	↑ 1.1	↑ 1.8	↓ 0.7	↓ 0.01	↑ 0.03

Visual Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
3D-CNN	56.1	58.4	24.9	1.31	0.26
CNN-LSTM	60.7	61.2	25.1	1.27	0.30
LSTM-FA	62.1	63.7	26.2	1.23	0.33
CMKL-V	52.6	58.5	29.3	-	-
SAL-CNN-V	63.8	-	-	-	-
SVM-MD-V	59.2	60.1	25.6	1.24	0.36
TFN_{visual}	**69.4**	**71.4**	**31.0**	**1.12**	**0.50**
Δ^{SOTA}_{visual}	↑ 5.6	↑ 7.7	↑ 1.7	↓ 0.11	↑ 0.14

Acoustic Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
HL-RNN	63.4	64.2	25.9	**1.21**	0.34
Adieu-Net	59.2	60.6	25.1	1.29	0.31
SER-LSTM	55.4	56.1	24.2	1.36	0.23
CMKL-A	52.6	58.5	**29.1**	-	-
SAL-CNN-A	62.1	-	-	-	-
SVM-MD-A	56.3	58.0	24.6	1.29	0.28
$\text{TFN}_{acoustic}$	**65.1**	**67.3**	27.5	1.23	**0.36**
$\Delta^{SOTA}_{acoustic}$	↑ 1.7	↑ 3.1	↓ 1.6	↑ 0.02	↑ 0.02

4 Result

#	Spoken words + acoustic and visual behaviors	TFN-Acoustic	TFN-Visual	TFN-Language	TFN-Early	TFN	Ground Truth
1	"You can't even tell funny jokes" + frowning expression	-0.375	-1.760	-0.558	-0.839	-1.661	-1.800
2	"I gave it a B" + smile expression + excited voice	1.967	1.245	0.438	0.467	1.215	1.400
3	"But I must say those are some pretty big shoes to fill so I thought maybe it has a chance" + headshake	-0.378	-1.034	1.734	1.385	0.608	0.400
4	"The only actor who can really sell their lines is Erin Eckart" + frown + low-energy voice	-0.970	-0.716	0.175	-0.031	-0.825	-1.000

Reference

- https://vimeo.com/238236114
- https://arxiv.org/pdf/1707.07250.pdf
- https://www.aclweb.org/anthology/D18-1382.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9131698&tag=1