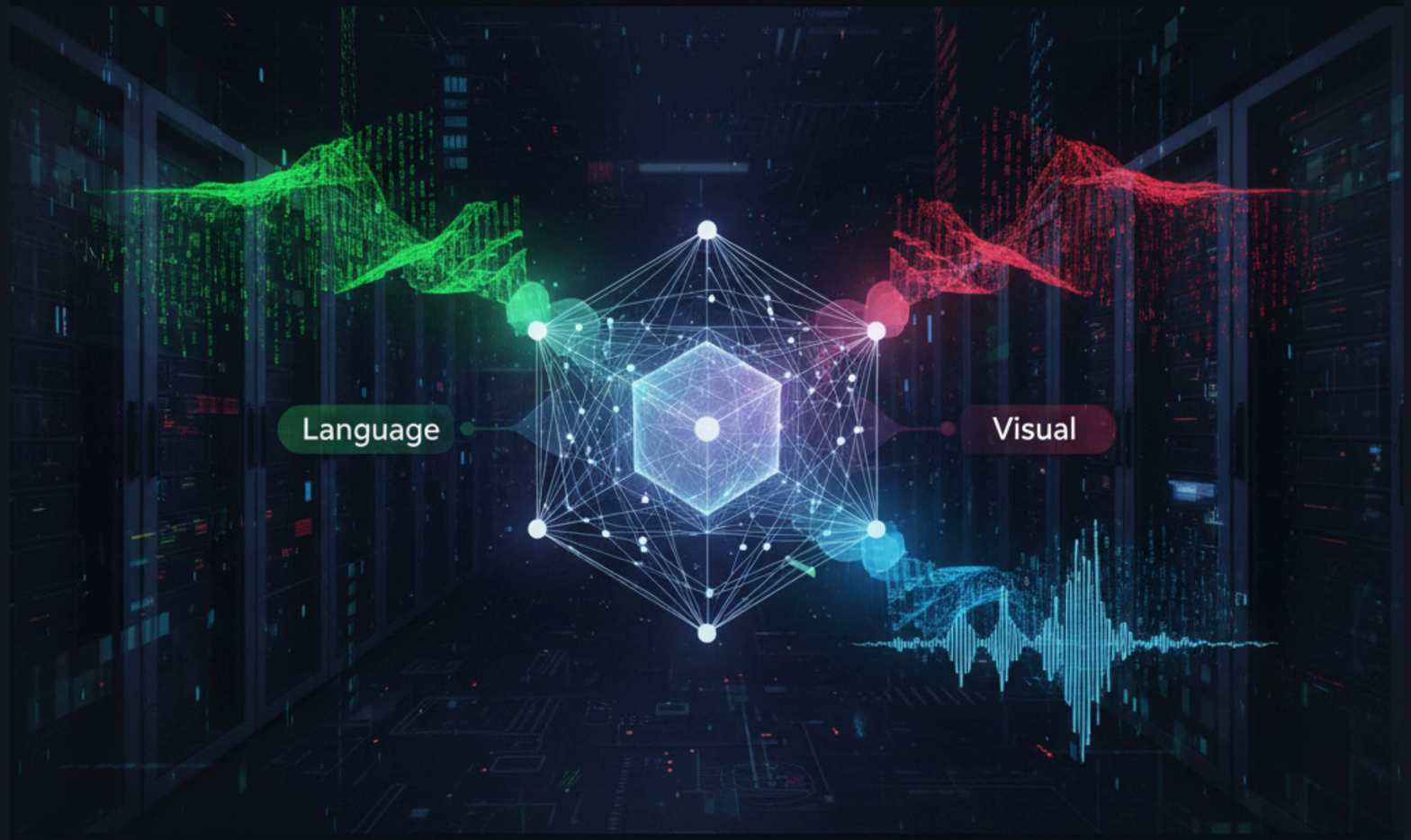
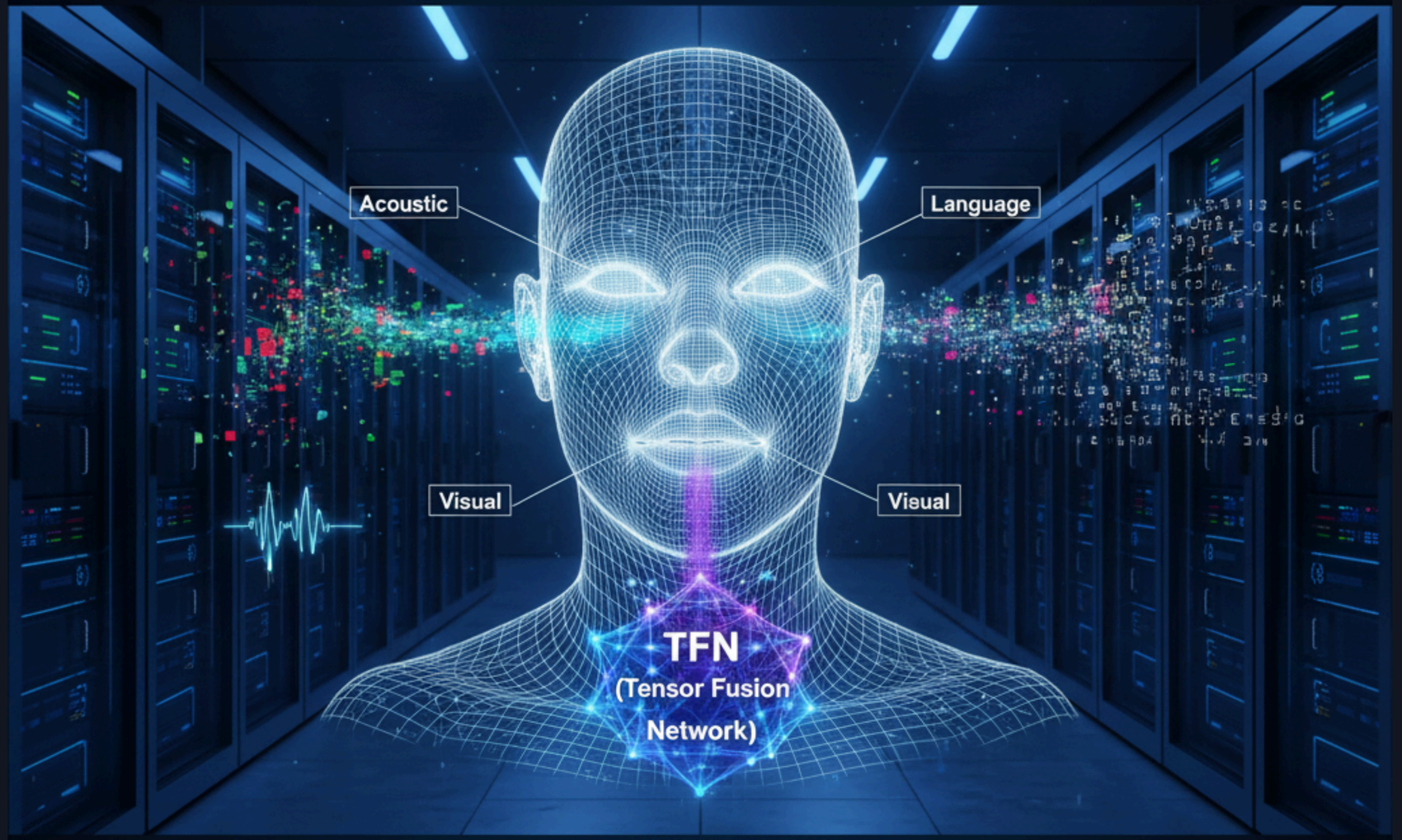
Tensor Fusion Network for Multimodal Sentiment Analysis



Amir Zadeh, Minghai Chen, Louis-Phlippe Morency
(Language Technologies Intitute, CMU)
Erik Cambria, Soujanya Poria

Introduction & Abstract



Problem: MSA requires modeling Intra-modality & Inter-modality dynamics.

Solution: Proposed TFN, an end-etod model to explicitty learn both dynamics.

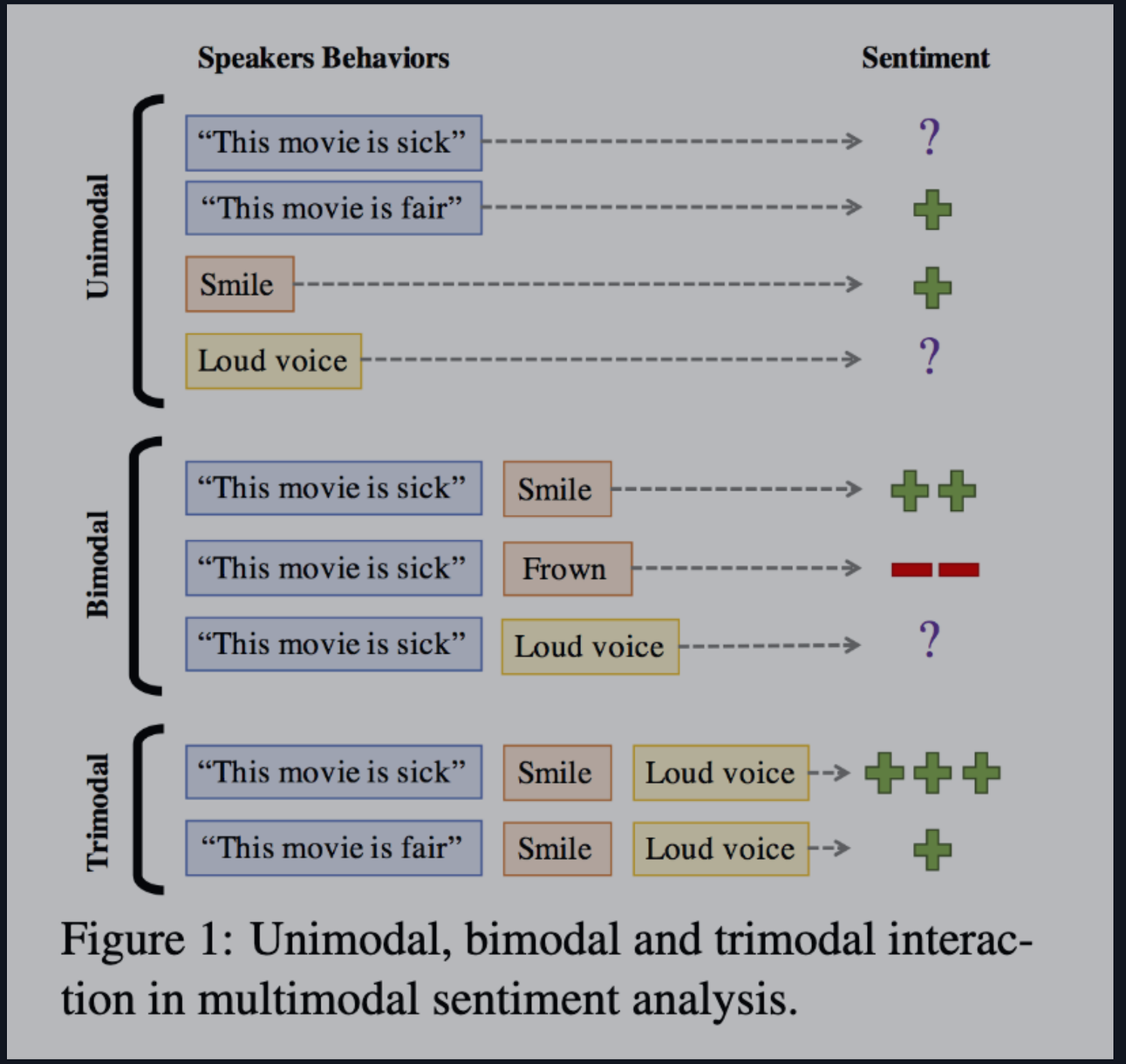
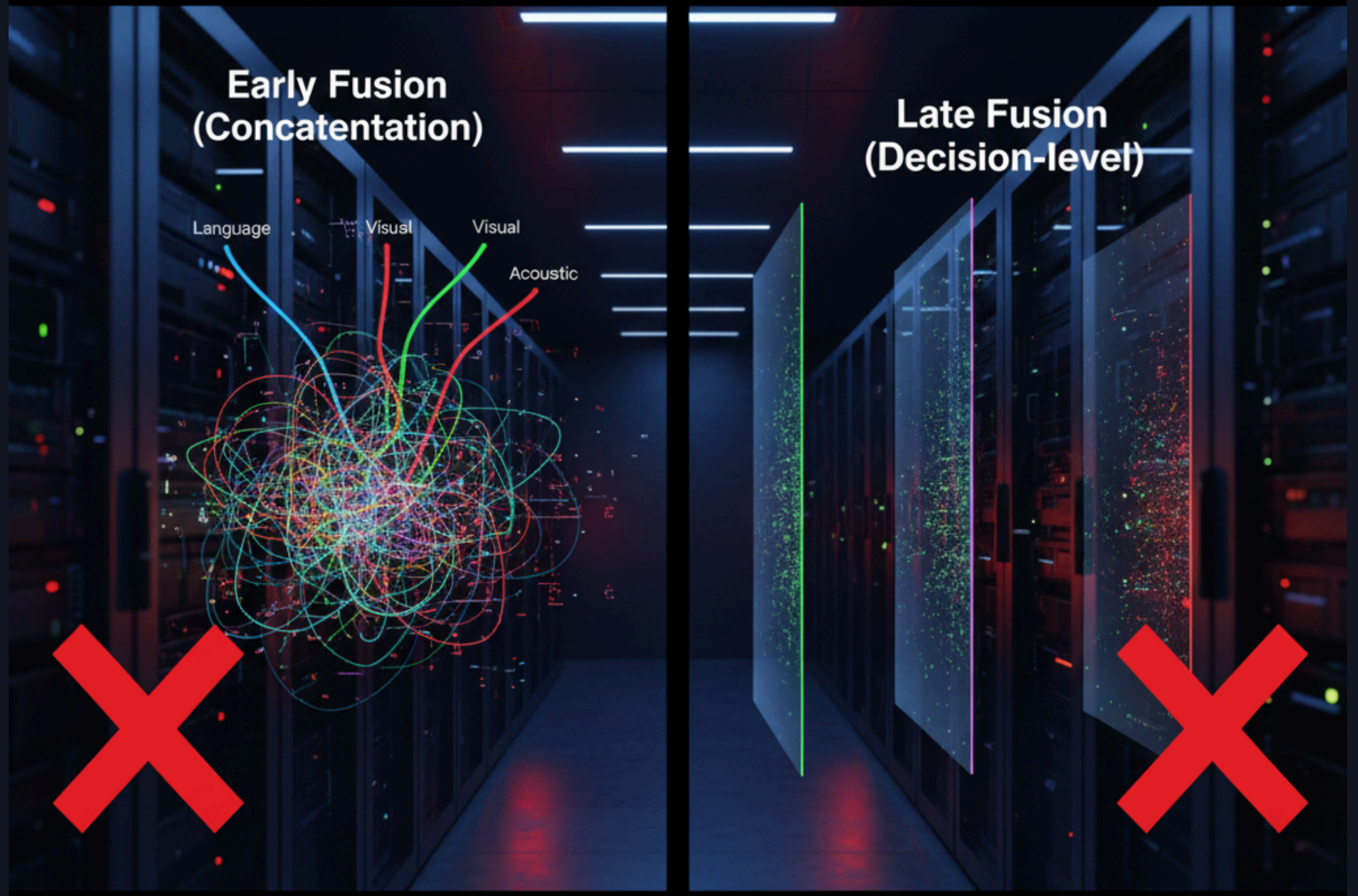


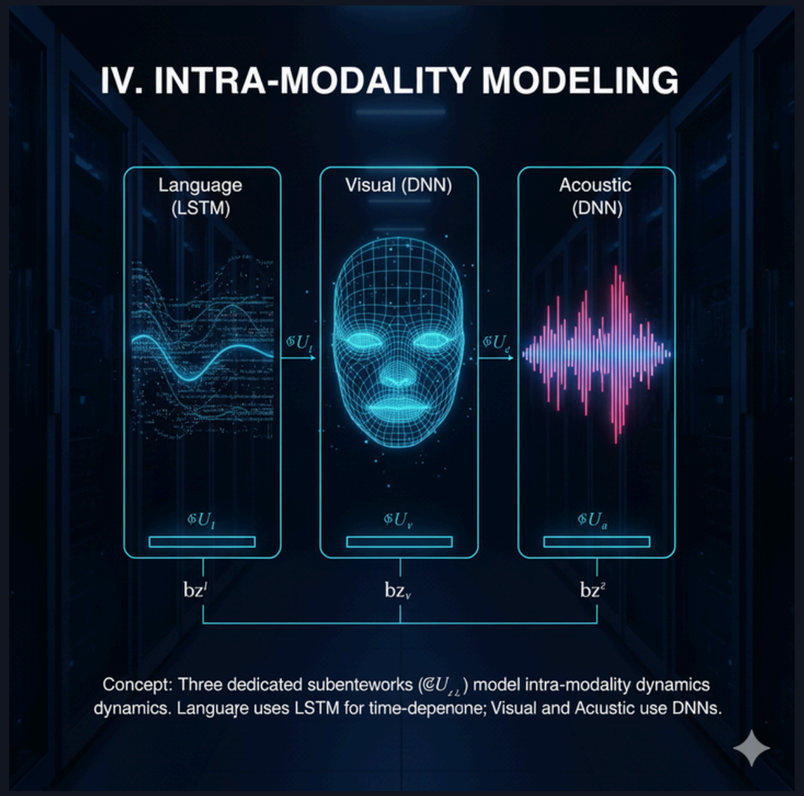
Figure 1: Unimodal, bimodal and trimodal interaction in multimodal sentiment analysis.

Limitations of Prior Fusion Methods



Problem: Previoues methods fail to capture necessary dynamics.
Early Fusion prevents effo-efficient Intra-modality modeling.
Late Fusion prevents learning crucial Inter-modality dynamics

TFN Architecture:
MES & Tensor Fusion & Inference



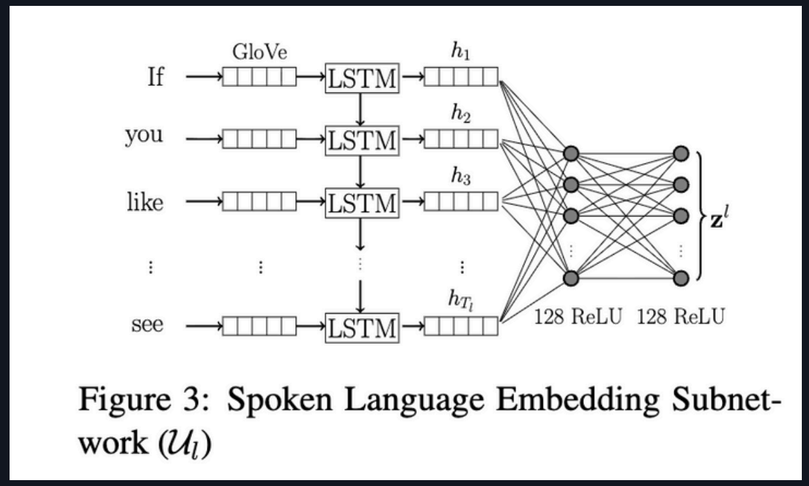


Figure 3: Spoken Language Embedding Subnetwork ($\mathcal{U}_l$)

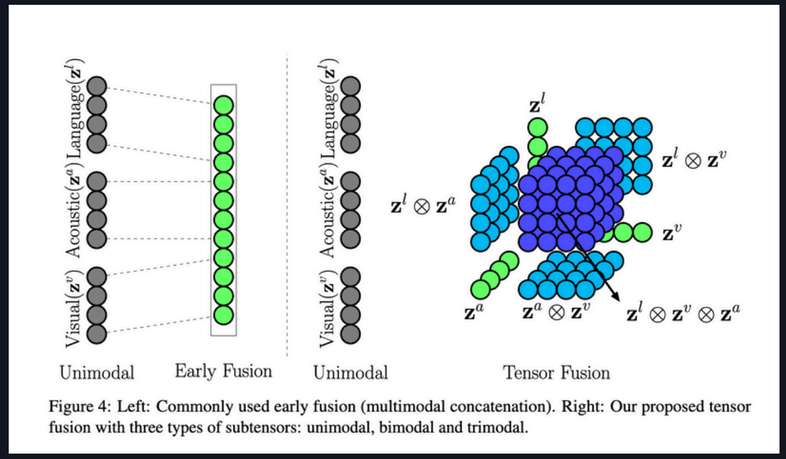


Figure 4: Left: Commonly used early fusion (multimodal concatenation). Right: Our proposed tensor fusion with three types of subtensors: unimodal, bimodal and trimodal.

Experimental Results

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

Table 1: Comparison with state-of-the-art approaches for multimodal sentiment analysis. TFN outperforms both neural and non-neural approaches as shown by Δ^{SOTA}.

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

Table 2: Comparison of TFN with its subtensor variants. All the unimodal, bimodal and trimodal subtensors are important. TFN also outperforms early fusion.

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

Table 3: Language Sentiment Analysis. Comparison of with state-of-the-art approaches for language sentiment analysis. $\Delta^{SOTA}_{language}$ shows improvement.

Experimental Results

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_{visual}^{SOTA}	↑ 5.6	↑ 7.7	↑ 1.7	↓ 0.11	↑ 0.14

Table 4: Visual Sentiment Analysis. Comparison with state-of-the-art approaches for visual sentiment analysis and emotion recognition. Δ_{visual}^{SOTA} shows the improvement.

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**
$\Delta_{acoustic}^{SOTA}$	↑ 1.7	↑ 3.1	↓ 1.6	↑ 0.02	↑ 0.02

Table 5: Acoustic Sentiment Analysis. Comparison with state-of-the-art approaches for audio sentiment analysis and emotion recognition. $\Delta_{acoustic}^{SOTA}$ shows improvement.

Conclusion & Future Work

- Novel Model: Introduced the Tensor Fusion Network (TFN), an end-to-end framework for Multimodal Sentiment Analysis.
- Explicit Fusion: Proposed the Tensor Fusion Layer, which uses a 3-fold Cartesian product to explicitly model unimodal, bimodal, and trimodal interactions.
- Custom Subnetworks: Developed dedicated Modality Embedding Subnetworks tailored for the characteristics of each modality in online video data (e.g., LSTM for volatile spoken language).
- Benchmark Results: Achieved new state-of-the-art performance on the CMU-MOSI dataset.

Thank you