MULTIMODAL SENTIMENT ANALYSIS BASED ON 3D STEREOSCOPIC ATTENTION

Jian Huang*, YuanYuan Pu*,☆, Dongming Zhou*, Hang Shi*, ZhengPeng Zhao*, Dan Xu*, Jinde Cao◊

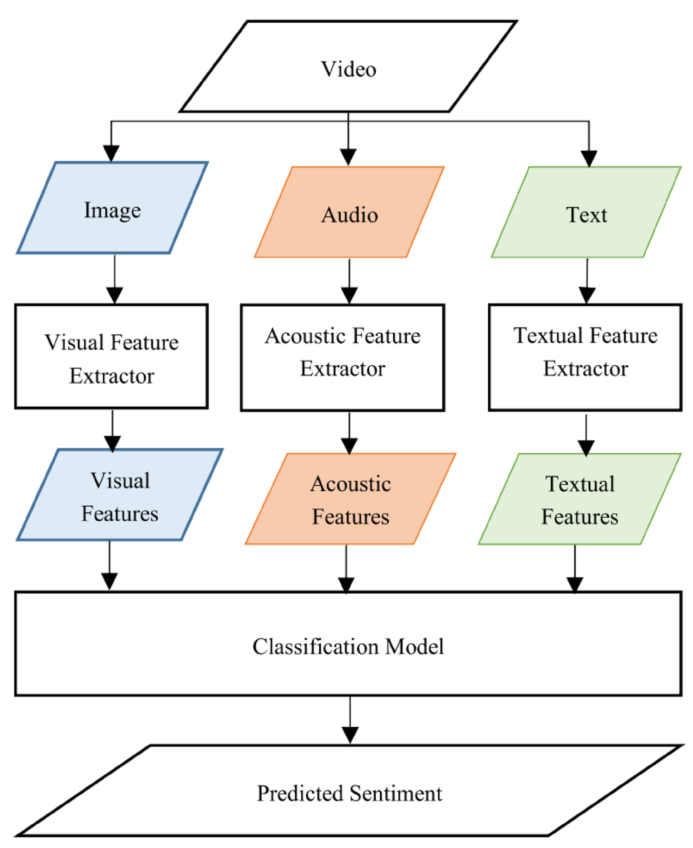
*YunNan University , School of Information Science and Engineering

☆University Key Laboratory of Internet of Things Technology and ApplicationYunnan Province

◊Southeast University, China and Yonsei Frontier Lab, Yonsei University, South Korea

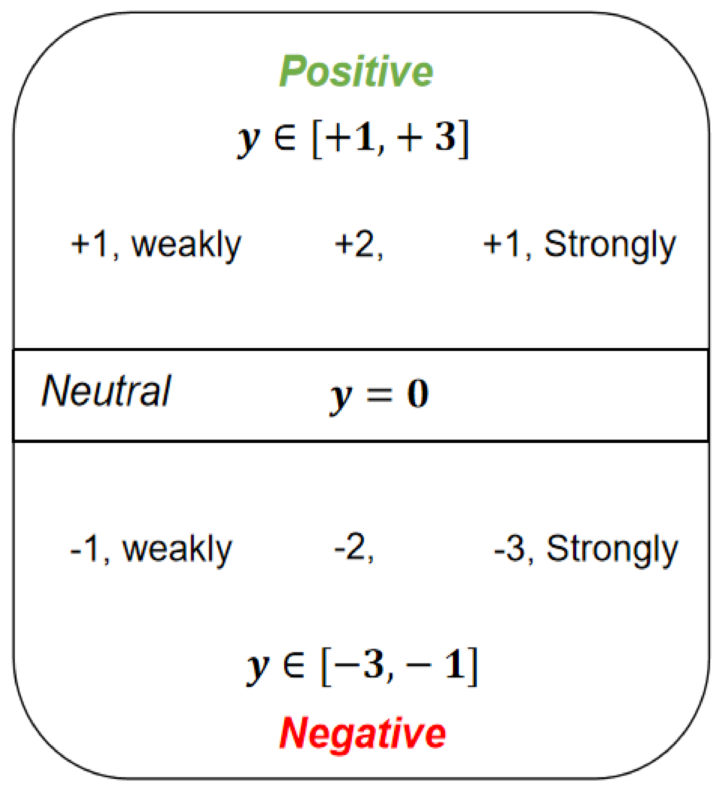
Background

Multimodal Sentiment Analysis



Sentiment Polarity

$y \in [-3, +3]$



Motivation

eg. a smile coupled with a positive word is positive, while audio represents sarcasm and ultimately leads to an opposite shift to negative.

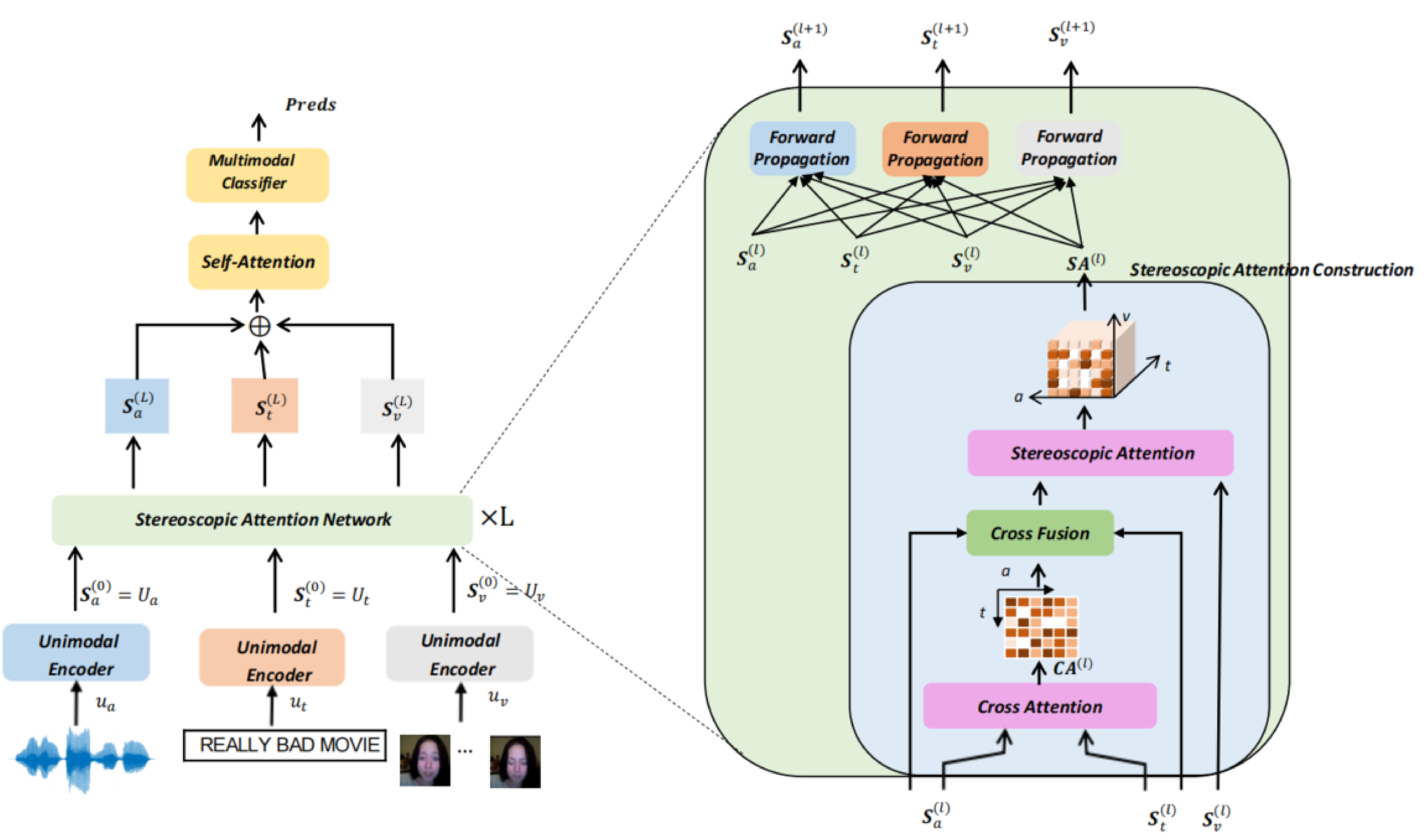
Problems with the current mainstream approach:

- RNNs only integrate unimodal information and miss interactions between modalities;
- The attention mechanism of the Transformer only explores the bi-modal interactions;
- General graphs contain only two nodes per edge.

Key challenges:

- **Different from 2D attention, there is dimension rising in the 3D attention generation;**
- **Based on the generated 3D attention, how to integrate the information between each 2D sequence modality**

Overall framework



Challenge 1 → Progressive Stereoscopic Attention

Take TA,V as an example:

Firstly, 2D cross-modal attention[3] is generated：

$$CA^{(l)} = Softmax\left(\frac{W_t^{(l)} S_t^{(l)} W_a^{(l)\mathrm{T}} S_a^{(l)\mathrm{T}}}{\sqrt{d}}\right)$$

Then, the 2D cross-modal attention weights are implemented to generate 3D attention：

$$F_{i,j}^{(l)} = S_{t,i}^{(l)} + CA_{i,j}^{(l)} \odot S_{a,j}^{(l)}, \{i \in T_t, j \in T_a\} \quad SA^{(l)} = Softmax\left(\frac{F^{(l)} W_v^{(l)\mathrm{T}} S_v^{(l)\mathrm{T}}}{\sqrt{d}}\right)$$

Challenge 2 → The forward propagation for stereoscopic attention

Taking the text modality as an example：

$$S_{v\to t}^{(l)} = \sum_{i=0}^{T_a} \frac{\left(SA^{(l)} S_v^{(l)}\right)_{T_t \times i \times d}}{T_a}$$

The text features are modulated by the audio and visual information, and the modulation factor α and β ensure the offset within a reasonable range.

$$S_{a\to t}^{(l)} = \sum_{j=0}^{T_v} \frac{\left(SA^{(l)} S_a^{(l)}\right)_{T_t \times j \times d}}{T_v}$$

$$S_t^{(l+1)} = S_t^{(l)} + \alpha * S_{v\to t}^{(l)} + \beta * S_{a\to t}^{(l)}$$

Experiments and results

1. Results on CMU-MOSI and CMU-MOSEI datasets

Model	CMU-MOSI				CMU-MOSEI			
	Acc2↑	F1↑	MAE↓	Corr↑	Acc2↑	F1↑	MAE↓	Corr↑
MFN[2]	77.26	77.38	0.9534	0.6672	80.23	80.77	0.5693	0.7202
MulT[3]	81.47	81.4	0.7892	0.7763	80.90	80.87	0.5690	0.7240
MISA[8]	81.95	81.91	0.7596	0.7771	81.10	81.18	0.5710	0.7310
BERT-MAG[9]	82.42	82.38	0.7313	0.7836	81.90	82.32	0.5640	0.7597
BBFN[5]	80.33	80.32	0.8216	0.6277	82.29	82.21	0.5820	0.7270
Self-MM[10]	82.51	82.47	0.7251	0.7896	82.17	82.46	0.5351	0.7605
MMIM[11]	82.33	82.28	0.7514	0.7718	80.04	80.47	0.5794	0.7352
UEGD*[12]	79.90	79.90	0.8860	0.6910	81.20	81.70	0.5430	0.7480
DMD[13]	82.07	82.08	0.7470	0.7859	82.81	83.06	0.5473	0.7518
Ours	**83.68**	**83.71**	**0.7124**	**0.7941**	**83.87**	**83.91**	**0.5210**	**0.7680**

2.Ablation Study

Table 2. Ablation study of our proposed method on CMU-MOSI. "w/o" denotes removing the component.

Model	Acc2↑	F1↑	MAE↓	Corr↑
w/o SA	80.32	80.21	0.9012	0.7351
UniAtten	81.25	81.29	0.8797	0.7418
CrossAtten	82.93	82.96	0.8452	0.7492
TV,A	83.23	82.87	0.7194	0.7585
AV,T	83.54	83.28	0.7185	0.7691
1 Layer	82.05	82.02	0.7430	0.7679
3 Layer	82.16	82.14	0.7342	0.7705
4 Layer	81.40	81.47	0.7496	0.7660
Ours	**83.68**	**83.71**	**0.7124**	**0.7941**

3.Case Study