

# DENSE FUSION NETWORK WITH MULTIMODAL RESIDUAL FOR SENTIMENT CLASSIFICATION

Huan Deng<sup>1</sup>, Peipei Kang<sup>1</sup>, Zhenguo Yang<sup>1,✉</sup>, Tianyong Hao<sup>2</sup>, Qing Li<sup>3</sup>, Wenyin Liu<sup>1,4,✉</sup>

<sup>1</sup>Guangdong University of Technology, Guangzhou, China

<sup>2</sup>South China Normal University, Guangzhou, China

<sup>3</sup>The Hong Kong Polytechnic University, Hong Kong, China

<sup>4</sup>Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen, China

d\_huan@163.com, ppkanggdut@126.com, yzg@gdut.edu.cn,  
haoty@126.com, qing-prof.li@polyu.edu.hk, liuwuy@gdut.edu.cn

## ABSTRACT

In this paper, we propose a deep dense fusion network with multimodal residual (DFMR) to integrate multimodal information including language, acoustic speeches, and visual images for sentiment analysis. DFMR exploits a dense fusion (DF) block to fuse the multimodal features obtained by modality-specific sequence networks, which is achieved by modelling their unimodal, bimodal and trimodal interactions jointly. Instead of concatenating the multimodal features directly, DF block conducts fusion for any two paired modalities firstly, and the fused information will be integrated with the other modalities subsequently. Furthermore, DFMR stacks multiple DF blocks to capture high-level semantic information conveyed by the multimodal representations. In particular, DFMR adopts a multimodal residual (MR) block to integrate the modality-specific features and fused features in each DF blocks, to avoid forgetting the multi-aspect information and alleviate gradient vanishing during stacking. Extensive experiments conducted on four public benchmark datasets show that DFMR outperforms eleven state-of-the-art baselines.

**Index Terms**— Multimodal Sentiment Analysis

## 1. INTRODUCTION

Social media platforms (e.g., YouTube, Twitter and Facebook, etc.) make people convenient to express opinions and emotions via sharing videos, images, audio, text, etc. The huge amount and public multimodal data attract a lot of research attention on sentiment analysis.

In the context of sentiment analysis, quite a few works have been proposed to exploit multimodal data. For instance, Williams et al. [1] concatenate the inputs from different modalities at each time-step and use that as the input to a single LSTM, which neglects cross-modal interactions. Zadeh

et al. [2] apply outer product on unimodal feature vectors to learn multimodal representations, which cannot model a long sequence through average temporal features for each modality. Zadeh et al. [3] use multi-attention blocks to capture the information across the three modalities, which cannot model cross-modal interactions in multiple levels.

In this paper, we design a deep dense fusion network with multimodal residual (DFMR) to learn modality-specific and cross-modal dynamics. DFMR firstly encodes unimodal independently by using a Bidirectional Gated Recurrent Unit (Bi-GRU) to model modality-specific interactions. Furthermore, DFMR adopts dense fusion (DF) blocks to model cross-modal interactions by fusing feature vectors for two paired modalities. We interpret multimodal fusion as a hierarchical interactive learning process, in which bimodal interactions are first learned based on unimodal dynamics, and then trimodal dynamics are achieved based on bimodal dynamics. In particular, DFMR adopts a multimodal residual (MR) module to retain the fused features of each layer, especially the ones extracted by the early layers. Finally, the feature representation achieved by MR can be sent to the sentiment classification module for predictions.

In summary, the main contributions are listed below.

- We propose an end-to-end deep dense fusion network with multimodal residual (DFMR) to fuse the multimodal features in a stacking manner for sentiment analysis.
- We devise a dense fusion (DF) block to fuse the modality-specific features with the fused features in a dense manner, modeling unimodal, bimodal and trimodal interactions simultaneously.
- We design a multimodal residual (MR) block to integrate the modality-specific features and fused features in each DF blocks, which keeps the low-level information in the deep networks.

✉ Corresponding authors

- We conduct extensive experiments on four benchmark datasets, which shows that the proposed DFMR achieves the state-of-the-art performance.

The paper is organized as follows. Section 2 reviews the related works. The details of the proposed network framework DFMR are introduced in Section 3. Experimental results and discussions are shown in Section 4, followed by conclusion in Section 5.

## 2. RELATED WORK

In the context of sentiment analysis, multimodal data usually are exploited and quite a few fusion methods have been proposed. For instance, some works [4] use concatenated feature vectors as input to make predictions or learn different models for each modality and combine the outputs using decision voting [5] without cross-modal interactions. To model modality-specific and cross-modal interactions, Zadeh et al. propose Tensor Fusion Network [2] through conducting outer product explicitly to aggregate unimodal, bimodal and trimodal interactions, followed by Low-rank Multimodal Fusion (LMF) [6]. These methods average temporal features for each modality to obtain the representation, which are difficult to model long sequences, because the average statistics cannot correctly capture contextual information. Zadeh et al. propose Memory Fusion Network (MFN) [7], which uses self-attention as a fusion method and remembers LSTM information through memory slot overtime. Wang et al. [8] propose Recurrent Attended Variation Embedding Network (RAVEN), which considers the fine-grained structure of non-verbal sub-word sequences and dynamically shifts the word representations based on nonverbal behaviors. The aforementioned methods model cross-modal interactions in single layer, which cannot capture multi-level interactions among the modalities.

## 3. METHODOLOGY

The framework of DFMR is shown in Fig. 1, which consists of four modules, i.e., the modality-specific module, dense multimodal fusion module, multimodal residual module, and the sentiment classification module. Given a piece of multimodal sequential data, including language, acoustic, and visual modality. Its modality-specific features are firstly extracted through the modality-specific module. Furthermore, the dense multimodal fusion module fuses the multimodal information from the modality-specific features with dense fusion (DF) blocks, and achieves fused multimodal features. Meanwhile, the fused features are sent into the multimodal residual (MR) module to generate multimodal residual features for sentiment classification.

### 3.1. Notation

In this work, we focus on tri-modality sequential data with respect of language ( $l$ ), acoustic ( $a$ ), and visual ( $v$ ) modalities. Denote a piece of multimodal sequential data as  $X = \{X^m\}$ ,  $m \in \{l, a, v\}$ . Each modality of the sequential data is further represented as  $X^m = [x_1^m, \dots, x_t^m, \dots, x_T^m] \in R^{T \times d_m}$ , where  $T$  is the sequence length,  $x_t^m$  is the  $t$ -th unit such as a word or an image, and  $d_m$  is the initial dimension of the unit. With the modality-specific module, we can get the modality-specific feature  $Z^m \in R^{d_v}$ , where  $d_v$  is the new dimension of the sequence. Taking  $Z^m$  as the initial state of the dense multimodal fusion module, the multimodal fused features can be obtained as  $F_i$ ,  $i = 1, 2, \dots, c$ , where  $c$  is the number of network blocks in this module. The residual features are denoted as  $R_i$ ,  $i = 1, 2, \dots, c$ , and the last residual feature  $R_i$  is fed into a sentiment classifier. Let  $Y$  denote the label of  $X$ , and  $K$  represent the number of all categories.

### 3.2. Modality-specific Module

As shown in Fig. 1, we introduce bi-directional gated recurrent unit ( $Bi - GRU_m$ ) to model time-aware interactions among the data units for each sequential modality. Furthermore, the modality-specific layers ( $MSL$ ) are designed to model modality-specific interactions. More specifically, each modality of the sequential data  $X^m$  is sent to  $Bi - GRU_m : R^{T \times d_m} \rightarrow R^{T \times d_{lat}}$  to learn context information, where  $Bi - GRU_m$  represents  $Bi - GRU$  for the modality  $m$ , and  $d_{lat}$  is the latent dimension of each unit in  $X^m$ . The process can be formulated as:

$$H^m = Bi - GRU_m(X^m, \Theta_{GRU}^m) \quad (1)$$

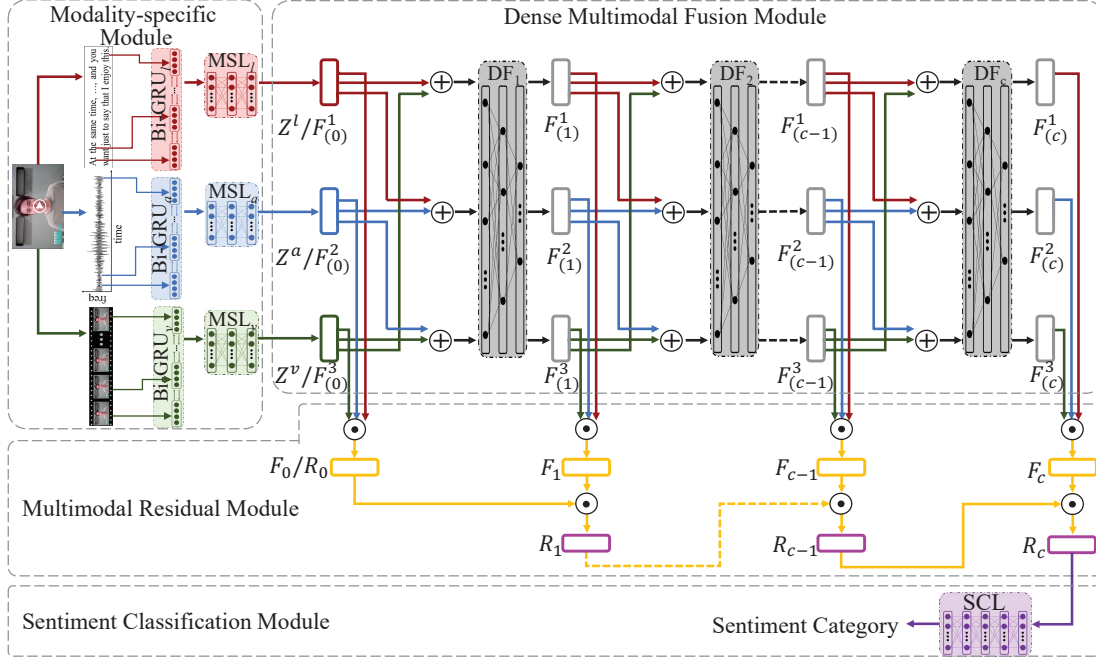
where  $H^m = [h_1^m, \dots, h_t^m, \dots, h_T^m] \in R^{T \times d_{lat}}$  denotes the latent sequential features in modality  $m$ ,  $h_t^m$  is the latent feature of the  $t$ -th unit, and  $\Theta_{GRU}^m$  are the parameters of  $Bi - GRU_m$ . To further explore the modality-specific interactions,  $MSL_m : R^{T \times d_{lat}} \rightarrow R^{d_v}$  is introduced with two fully-connected layers activated by  $ReLU$  for modality  $m$ . By concatenating all the unit features as input, the modality-specific features can be obtained as below,

$$Z^m = MSL_m \left( Cat \left( h_t^m \Big|_{t=1}^T \right), \Theta_{VS}^m \right) \quad (2)$$

where  $Cat(\cdot)$  denotes the concatenation operation, and  $\Theta_{VS}^m$  are the parameters of  $MSL_m$ .

### 3.3. Dense Multimodal Fusion Module

Given the modality-specific features of the multimodal sequential data, we devise a dense multimodal fusion network to conduct multimodal fusion. As shown in Fig. 1, each two modality-specific features are fused together in the first fusion block to model bimodal interactions. For the subsequent blocks, each two fused multimodal features are further fused



**Fig. 1.** The framework of DFMR. Note that  $\oplus$  denotes the concatenation operation, and  $\odot$  denotes the sum operation.

to integrate multi-aspects information. Assume that  $F_{(i)}^j$  is the  $j$ -th fused multimodal feature in the  $i$ -th block, where  $j \in 1, 2, 3, i \in 1, 2, \dots, c$ , and  $c$  is the number of blocks,  $F_{(i)}^j$  is achieved by fusing two former fused features of the  $(i-1)$ -th block through dense fusion (DF) blocks. In particular, a DF block is composed of two fully-connected layers activated by *ReLU*, which is formulated as,

$$F_{(i)}^j = \begin{cases} DF_i \left( \text{Cat} \left( F_{(i-1)}^j, F_{(i-1)}^{j+2} \right), \Theta_{DF}^i \right), j=1 \\ DF_i \left( \text{Cat} \left( F_{(i-1)}^j, F_{(i-1)}^{j-1} \right), \Theta_{DF}^i \right), j=2, 3 \end{cases} \quad (3)$$

where  $DF_i : \mathbb{R}^{2d_v} \rightarrow \mathbb{R}^{d_v}$  denotes the  $i$ -th DF block,  $\Theta_{DF}^i$  are the corresponding parameters, and  $\text{Cat}(\cdot)$  is the concatenation operation. In order to describe the whole dense fusion module conveniently, we rewrite the modality-specific features  $Z^m$  as the fused multimodal features of the 0-th block, i.e., the initialization of the dense multimodal fusion module:

$$F_{(0)}^1 = Z^l, F_{(0)}^2 = Z^a, F_{(0)}^3 = Z^v \quad (4)$$

Consequently, deeper multimodal information can be fused from the modality-specific features, and the inter-modality interaction is explored.

### 3.4. Multimodal Residual Module

Given the fused features achieved by stacking DF blocks, we propose multimodal residual (MR) block to transfer the fused multimodal information from the former blocks to the latter

blocks, as shown in Fig. 1. As a result, MR block is able to remember the fused features that are achieved from DF blocks, and alleviate gradient vanishing of deep models.

Specifically, for the  $i$ -th multimodal residual block, we expect to obtain a residual feature  $R_i \in \mathbb{R}^{d_v}, i \in \{1, 2, \dots, c\}$ , which is supposed to remember the current multimodal information from the  $i$ -th multimodal fusion block, and also the former multimodal residual information from the  $(i-1)$ -th block. It is formulated as:

$$R_i = F_i + R_{(i-1)} \quad (5)$$

where  $F_i \in \mathbb{R}^{d_v}, i \in \{0, 1, \dots, c\}$  denotes the multimodal information from the  $i$ -th multimodal fusion block, and it is achieved by adding all the fused multimodal features in this block, as shown in Eq. (6),

$$F_i = \sum_{j=1}^3 F_{(i)}^j \quad (6)$$

For convenient description, we rewrite  $F_0$  as the residual feature of the 0-th multimodal residual block, i.e., the initialization of the multimodal residual module:  $R_0 = F_0$ . By continuously stacking the multimodal fusion blocks and the multimodal residual blocks, we finally obtain the global feature representation  $R_c$  from the last block that records the hierarchical multimodal fusion information.

### 3.5. Sentiment Classification Module

Given the fused features achieved by DF and MR blocks, we devise the sentiment classification layers ( $SCL : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^K$ ) with four fully-connected layers activated by *ReLU* for sentiment classification, where  $K$  is the number of categories. The objective function can be defined as the difference between the true label  $Y$  and the predicted result:

$$obj = G(Y, SCL(R_c, \Theta_{SCL})) \quad (7)$$

where  $\Theta_{SCL}$  are the parameters of  $SCL$ , and  $G(\cdot)$  represents the mean square error loss or the cross entropy loss.

## 4. EXPERIMENTS

### 4.1. Datasets

There are four public datasets for evaluations. 1) CMU-MOSI [9] consists of 93 YouTube videos, which are divided into 2,199 multimodal opinion segments, consisting of languages, speeches, and videos. Each opinion segment is annotated from -3 to 3. 2) ICT-MMMO [10] collects 370 review videos, and each review video is divided into three modalities and annotated with negative and positive. 3) YouTube [11] collects 47 videos covering topics, such as job, business, and war, etc. These videos are divided into 269 multimodal segments and annotated as negative, neutral, or positive, respectively. 4) IEMOCAP [12] contains 151 recorded video dialogues with 302 videos, which are divided into 10k multimodal utterances. We take 7,318 utterances with labels of happy, sad, angry and neutral by following [8] [13].

For evaluations, binary accuracy (Acc-2), 3-class accuracy (Acc-3), 7-class accuracy (Acc-7),  $F_1$  score, mean absolute error (MAE), correlation (Corr) are adopted.

### 4.2. Baselines

The baselines include a number of sentiment analysis approaches: Tensor Fusion Network (TFN) [2], Low rank Multimodal Fusion (LMF) [6], Memory Fusion Network (MFN) [7], Dynamic fusion graph (DFG) [14], Early Fusion LSTM (EF-LSTM) [1], Bi-directional Contextual LSTM (BC-LSTM) [15], Convolutional Multiple Kernel Learning (C-MKL) [5], Multi-attention Recurrent Network (MARN) [3], Recurrent Multistage Fusion Network (RMFN) [13], Multimodal Factorization Model (MFM) [16], Recurrent Attended Variation Embedding Network (RAVEN) [8].

### 4.3. Compared with the baselines

**Table 1.** Performance on ICT-MMMO and YouTube datasets.

Method	ICT-MMMO		YouTube	
	Acc-2	$F_1$	Acc-3	$F_1$
C-MKL	80.0	72.4	50.2	50.8
BC-LSTM	70.0	70.1	45.0	45.1
EF-LSTM	66.3	65.0	44.1	43.6
TFN	72.5	72.6	45.0	41.0
MARN	71.3	70.2	48.3	44.9
MFN	73.8	73.1	51.7	51.6
MFM	81.3	79.2	53.3	52.4
DFMR	<b>82.5</b>	<b>82.4</b>	<b>61.7</b>	<b>61.4</b>

**1) On ICT-MMMO and YouTube datasets.** The overall performance of the approaches are summarized in Table 1, from which we have some observations. Firstly, compared with the non-temporal models, e.g., TFN, DFMR performs better, benefiting from using Bi-GRU to learn the contextual and temporal information of utterances. Secondly, DFMR outperforms the temporal models including MARN, MFM, etc., benefiting from stacking DF blocks in multiple levels instead of just using single network layer.

**Table 2.** Performance on CMU-MOSI dataset.

Methods	Acc-2	Acc-7	$F_1$	MAE	Corr
C-MKL	72.3	30.2	72.0	-	-
MFN	77.4	34.1	77.3	0.965	0.632
DFG	77.7	35.6	77.7	0.96	0.66
BC-LSTM	73.9	28.7	73.9	1.079	0.581
EF-LSTM	74.3	32.4	74.3	1.023	0.622
TFN	74.6	28.7	74.5	1.040	0.587
MARN	77.1	34.7	77.0	0.968	0.625
RMFN	78.4	<b>38.3</b>	78.0	0.922	0.681
MFM	78.1	36.2	78.1	0.951	0.662
RAVEN	78.0	33.2	76.6	0.915	<b>0.691</b>
LMF	76.4	32.8	75.7	<b>0.912</b>	0.668
DFMR	<b>79.3</b>	30.3	<b>79.2</b>	0.961	0.663

**2) On CMU-MOSI and IEMOCAP datasets.** From Table 2 and Table 3, we can observe that DFMR achieves competitive performance. The imbalanced number of samples for the sentiment categories is the reason that makes DFMR focus more on the ones with sufficient samples. Overall, the experimental results show the effectiveness of DFMR.

**Table 3.** Performance on IEMOCAP dataset.

Method	Happy		Sad		Angry		Neutral	
	$F_1$	Acc-2	$F_1$	Acc-2	$F_1$	Acc-2	$F_1$	Acc-2
TFN	83.6	84.8	82.8	83.4	84.2	83.4	65.4	67.5
BC-LSTM	81.7	84.9	81.7	83.2	84.2	83.5	64.1	67.5
EF-LSTM	83.3	85.2	81.1	82.1	84.3	84.5	67.1	68.2
MARN	83.6	86.7	81.2	82.0	84.2	84.6	65.9	66.8
MFN	85.3	90.1	79.2	85.8	86.0	87.0	61.7	71.8
MFM	85.8	90.2	<b>86.1</b>	88.4	86.7	87.5	68.1	72.1
LMF	85.8	87.3	85.9	86.2	<b>89.0</b>	<b>89.0</b>	<b>71.7</b>	72.4
DFG	85.8	87.5	85.1	82.9	84.2	84.6	69.1	69.5
RAVEN	85.8	87.3	83.1	83.4	86.7	87.3	69.3	69.7
DFMR	<b>89.1</b>	<b>92.3</b>	84.7	<b>89.5</b>	81.7	87.5	70.6	<b>76.6</b>

**Table 4.** Impact of with or without (w/o) the MR block on the performance on CMU-MOSI dataset.

Method	Acc-2	Acc-7	$F_1$	MAE	Corr
w/o MR	45.77	15.45	34.72	1.421	0.061
DFMR	<b>79.30</b>	<b>30.32</b>	<b>79.17</b>	<b>0.961</b>	<b>0.663</b>

#### 4.4. Effectiveness of DF block

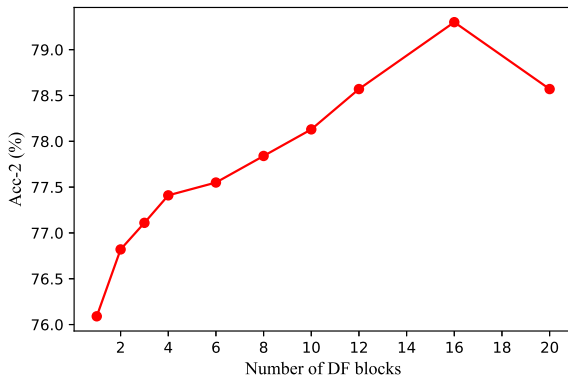
**Fig. 2.** Impact of the number of stacked DF blocks.

Fig. 2 summarizes the impact of the number of stacked DF blocks on CMU-MOSI dataset, from which we can observe that the performance tends to be improved gradually with the increase of the number of stacked DF blocks. However, too many DF blocks may increase the complexity of the networks, which is difficult to be trained.

#### 4.5. Effectiveness of MR block

To verify the effectiveness of the MR block, we take the CMU-MOSI dataset as an example, and report the perfor-

**Table 5.** Performance of DFMR using single modalities versus multiple modalities on CMU-MOSI dataset. Note that L denotes language modality, A denotes acoustic modality, and V represents visual modality.

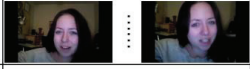

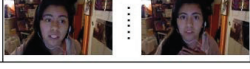

Method	Acc-2	Acc-7	$F_1$	MAE	Corr
L	76.53	27.41	76.32	1.029	0.600
A	57.00	17.20	56.45	1.404	0.156
V	60.35	18.08	60.45	1.378	0.181
L + A	77.26	29.16	76.90	1.004	0.638
L + V	75.80	29.45	75.86	1.037	0.628
A + V	59.77	16.18	55.04	1.383	0.208
L + A + V	<b>79.30</b>	<b>30.32</b>	<b>79.17</b>	<b>0.961</b>	<b>0.663</b>

mance of DFMR with or without MR block in Table 4. From the table, we can observe that the performance of DFMR on five metrics decrease dramatically without MR block. The reason is that with DFMR goes deeper, the low-level information extracted by the earlier layers may be forgotten, and the gradients may be faced with vanishing or exploding problems. The experimental results manifest the significance of the devised MR block.

#### 4.6. Effectiveness of multimodal fusion

The performance of DFMR using different data modalities are summarized in Table 5. In terms of the single modalities, language modality performs the best, as the semantic information conveyed by language is relatively explicit compared with acoustic and visual modalities. In particular, DFMR achieves significant improvement.

#### 4.7. Failure examples

#	Language	Visual	Acoustic	DFMR	Ground Truth
1	Oh my gosh bad movie			-1.0	-2.8
2	But in this oh my god i love you			1.2	2.5

**Fig. 3.** Failure examples on CMU-MOSI.

Fig. 3 shows some failure examples that the expressions of the modalities seem to be not consistent. For example, the sentiment of language in the first case is typically negative, while the images seem to be smiling faces, making the model confused. Though DFMR can recognize it as negative, it is evaluated as a totally wrong prediction for classification.

## 5. CONCLUSION

In this paper, we propose a deep dense fusion network with multimodal residual (DFMR) to integrate multimodal information including language, acoustic, and visual for sentiment analysis. DFMR stacks dense fusion (DF) blocks to obtain different levels of information, and exploit multimodal residual (MR) block to integrate the multi-level information and alleviate gradient vanishing in deep networks. DFMR has achieved significant performance compared with eleven state-of-the-art baselines on four benchmark datasets.

## 6. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No.62076073), the Guangdong Basic and Applied Basic Research Foundation (No.2020A1515010616), the Guangdong Innovative Research Team Program (No.2014ZT05G157), the Key-Area Research and Development Program of Guangdong Province (2019B010136001), and the Science and Technology Planning Project of Guangdong Province (LZC0023), and Hong Kong RGC CRF Project C1031-18G.

## 7. REFERENCES

- [1] Williams J., Kleinegesse S., Comanescu R., and Radu O., “Recognizing emotions in video using multimodal dnn feature fusion,” in *Challenge-HML*, pp. 11–19, 2018.
- [2] Zadeh A., Chen M., Poria S., Cambria E., and Morency L., “Tensor fusion network for multimodal sentiment analysis,” in *EMNLP*, 2017.

- [3] Zadeh A., Liang P., Poria S., Vij P., Cambria E., and Morency L., “Multi-attention recurrent network for human communication comprehension,” in *AAAI*, vol. 2018, pp. 5642, 2018.
- [4] Poria S., Chaturvedi I., Cambria E., and Hussain A., “Convolutional mkl based multimodal emotion recognition and sentiment analysis,” in *ICDM*, pp. 439–448, 2016.
- [5] Poria S., Cambria E., and Gelbukh A., “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” in *EMNLP*, pp. 2539–2544, 2015.
- [6] Liu Z., Shen Y., Lakshminarasimhan V., Liang P., Zadeh A., and Morency L., “Efficient low-rank multimodal fusion with modality-specific factors,” in *ACL*, 2018, pp. 2247–2256.
- [7] Zadeh A., Liang P., Mazumder N., Poria S., Cambria E., and Morency L., “Memory fusion network for multi-view sequential learning,” in *AAAI*, 2018, vol. 32.
- [8] Wang Y., Shen Y., Liu Z., Liang P., Zadeh A., and Morency L., “Words can shift: Dynamically adjusting word representations using nonverbal behaviors,” in *AAAI*, vol. 33, pp. 7216–7223, 2019.
- [9] Zadeh A., Zellers R., Pincus E., and Morency L., “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
- [10] Wöllmer M., Weninger F., Knaup T., Schuller B., Sun C., Sagae K., and Morency L., “Youtube movie reviews: Sentiment analysis in an audio-visual context,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.
- [11] Morency L., Mihalcea R., and Doshi P., “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in *ICMI*, pp. 169–176, 2011.
- [12] Busso C., Bulut M., Lee C., Kazemzadeh A., Mower E., Kim S., Chang J., Lee S., and Narayanan S., “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.
- [13] Liang P., Liu Z., Zadeh A., and Morency L., “Multimodal language analysis with recurrent multistage fusion,” in *EMNLP*, 2018.
- [14] Zadeh A., Liang P., Poria S., Cambria E., and Morency L., “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *ACL*, pp. 2236–2246, 2018.
- [15] Poria S., Cambria E., Hazarika D., Majumder N., Zadeh A., and Morency L., “Context-dependent sentiment analysis in user-generated videos,” in *ACL*, pp. 873–883, 2017.
- [16] Tsai Y., Liang P., Zadeh A., Morency L., and Salakhutdinov R., “Learning factorized multimodal representations,” in *ICLR*, 2019.