Hierarchical-gate Multimodal Network for Human Communication Comprehension

Qiyuan Liu, Liangqing Wu, Yang Xu, Dong Zhang*, Shoushan Li, and Guodong Zhou

School of Computer Science and Technology, Soochow University, Suzhou, China
{qyliu,lqwu,yxu2017,dzhang17}@stu.suda.edu.cn
{lishoushan,gdzhou}@suda.edu.cn

Abstract. Computational modeling of human multimodal language is an emerging research area in natural language processing spanning the language, visual and acoustic modalities. Comprehending multimodal language requires modeling not only the interactions within each modality (intra-modal interactions) but more importantly the interactions between modalities (cross-modal interactions). In this paper, we present a novel neural architecture for understanding human communication called the Hierarchical-gate Multimodal Network(HGMN). Specifically, each modality is first encoded by Bi-LSTM which aims to capture the intra-modal interactions within single modality. Subsequently, we merge the independent information of multi-modality using two gated layers. The first gate which is named as modality-gate will calculate the weight of each modality. And the other gate called temporal-gate will control each time-step contribution for final prediction. Finally, the max-pooling strategy is used to reduce the dimension of the multimodal representation, which will be fed to the prediction layer. We perform extensive comparisons on five publicly available datasets for multimodal sentiment analysis, emotion recognition and speaker trait recognition. HGMN shows state-of-the-art performance on all the datasets.

Keywords: Multimodal · Human communication · Hierarchical-gate.

1 Introduction

Computational modeling of human multimodal language is an upcoming research area in natural language processing. This research area focuses on modeling tasks such as multimodal sentiment analysis, emotion recognition, and personality traits recognition. We utilize three modalities to Communicate our intentions: language modality (words, phrases and sentences), vision modality (gestures and expressions), and acoustic modality (paralinguistics and changes in vocal tones). These multimodal signals are highly structured with two prime forms of interactions: intra-modal and cross-modal interactions [11]. Intra-modal interactions

* Corresponding Author.

represent information within a specific modality, independent of other modalities. Cross-modal interactions represent interactions between modalities. Modeling these interactions lies at the heart of human multimodal language analysis processing [2].

Intra-modal interactions are usually captured by Convolutional Neural Networks or Long Short-Term Memory Networks. The methods of getting cross-modal interactions are different. Traditional methods like TFN [14] use outer product to fuse different modalities which is a coarse-grained fusion method that can not capture the complex interactions between modalities. MARN(Multi-attention Recurrent Network) [16] utilizes attention mechanism to capture the importance between modalities, but ignores the temporal information. The existing methods do not take into account both modality and temporal information importances.

In order to overcome the challenges of the above methods, we propose a novel model called the Hierarchical-gate Multimodal Network(HGMN). Each modality is first encoded by Bi-LSTM which aims to capture the intra-modal interactions within single modality. Subsequently, we merge the independent information of multi-modality using two gated layers. The first gate which is named as modality-gate will calculate the weight of each modality. And the other gate called temporal-gate will capture the importances of temporal information. Finally, the max-pooling strategy is used to reduce the dimension of the multimodal representation, which will be fed to the prediction layer. We perform extensive comparisons on five publicly available datasets for multimodal sentiment, emotion analysis and speaker trait recognition. HGMN shows state-of-the-art performance on all the datasets.

2 Related Work

Researchers dealing with multimodal human communication have largely focused on three major types of models.

The first category is Early Fusion models which rely on concatenation of all modalities into a single view to simplify the learning setting. These approaches then use this concatenated view as input to a learning model. Hidden Markov Models (HMM) [1], Support Vector Machines (SVM) and Hidden Conditional Random Fields (HCRF) [10] have been successfully used for structured prediction.

The second category is Late Fusion models which learn different models for each modality and combine the outputs using decision voting [5, 12]. While these methods are generally strong in modeling intra-modal interactions, they have shortcomings for cross-modal interactions since these inter-modality interactions are normally more complex than a decision vote.

The third category of models rely on collapsing the time dimension from sequences by learning a temporal representation for each of the different modalities. Such methods have used average feature values over time [8]. Essentially these models apply conventional multi-modality learning approaches, such as

Multiple Kernel Learning, subspace learning or co-training to the Multimodal representations. Other approaches have trained different models for each view and combined the models using decision voting, tensor products or deep neural networks [9].

Different from the first category models, our proposed approach in this paper models both intra-modal and cross-modal interactions. In addition, different from the second and third categories, we simultaneously handle the modality contribution and sequence contribution in time-dependent interactions by two kinds of gated mechanisms.

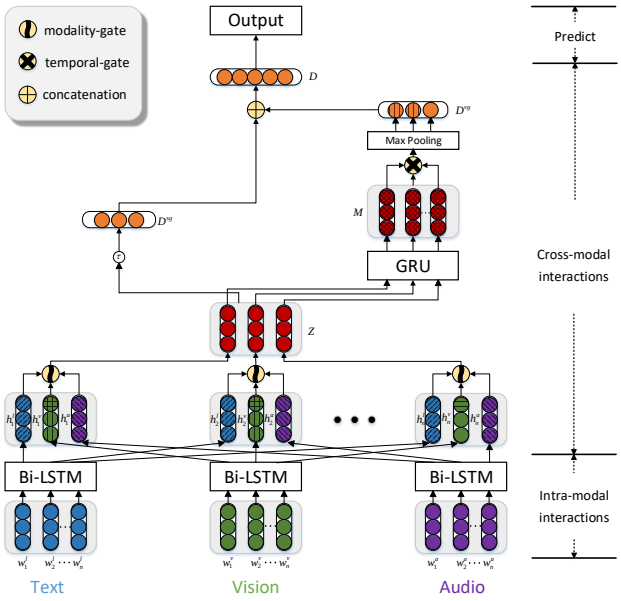

Fig. 1. Overview figure of Hierarchical-gate Multimodal Network(HGMN) pipeline.

3 HGMN Model

In this section we outline our pipeline for human communication comprehension: the Hierarchical-gate Multimodal Network(HGMN). Specifically, HGMN consists of three main components: 1) Intra-modal Interactions Calculation. 2) Cross-modal Interactions Identification which includes the Hierarchical-gate network. 3) Prediction layer. Figure 1 shows the overview of HGMN model.

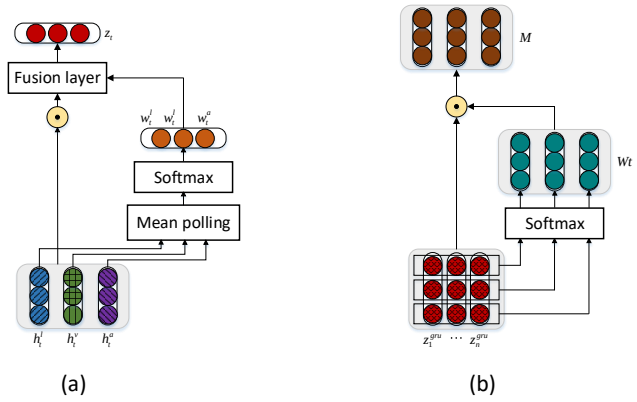

Fig. 2. Modality-Gate(a) and Temporal-Gate(b).

3.1 Intra-modal Interactions Calculation

The input to HGMN is a multi-modality sequence consisting of language, video, and audio for $M = \{l; v; a\}$. At first, in terms of language modality, we assume that an utterance contains n words. $w_t^l \in R^{d^l}$ represents the t-th word in the utterance. Then, we use a bidirectional LSTM (namely, Bi-LSTM) to encode the forward and backward contexts. The Bi-LSTM contains the forward $\overrightarrow{\text{LSTM}}$ which reads the utterance from w_1^l to w_n^l and a backward $\overleftarrow{\text{LSTM}}$ which reads from w_n^l to w_1^l:

$$\overrightarrow{h_t^l} = \overrightarrow{\text{LSTM}}(w_t^l, \overrightarrow{h_{t-1}^l}); t \in [1, n] \tag{1}$$

$$\overleftarrow{h_t^l} = \overleftarrow{\text{LSTM}}(w_t^l, \overleftarrow{h_{t+1}^l}); t \in [n, 1] \tag{2}$$

We obtain an annotation for a given word by concatenating the forward hidden state $\overrightarrow{h_t^l} \in R^d$ and backward hidden state $\overleftarrow{h_t^l} \in R^d$ as h_t^l, which summarizes the contextual information of whole utterance centered around the word w_t^l. The sequence output of the language modality is $H^l = [h_1^l; h_2^l; \cdots; h_n^l]; H^l \in R^{n \times 2d}$. Similarly, the vision modality and audio modality is represented as $H^v \in R^{n \times 2d}$ and $H^a \in R^{n \times 2d}$ after the individual Bi-LSTM over all time-steps.

3.2 Cross-modal Interactions Identification

In this subsection, we will introduce the cross-modal interactions identification. We have got the hidden state of each modality in last subsection. Let $h_t = [h_t^l; h_t^v; h_t^a]$ represent the modalities concatenated hidden state of the t-th word in the utterance ,$h_t \in R^{3 \times 2d}$.

Modality-Gate As the first gated layer, modality-gate will fuse different modalities according to the weight of each modality. Figure 2(a) shows the construction of the modality-gate. Modality-gate calculate the weight of each modality through two steps: 1) Calculate the mean of each modalitys hidden state. 2)Feed the concatenated mean of three modalities to a softmax layer.

$$a_t^m = \text{Meanpooling}(h_t^m); m \in \{l, v, a\} \tag{3}$$

$$[\mathrm{s}_t^l, s_t^v, s_t^a] = \text{softmax}([a_t^l, a_t^v, a_t^a]) \tag{4}$$

$$z_t = h_t^l \cdot s_t^l + h_t^v \cdot s_t^v + h_t^a \cdot s_t^a; z_t \in R^{2d} \tag{5}$$

s_t^m is the weight of modality m at time tth. Firstly, we multiply the features of each modality with corresponding weight. Then, summing the weighted modalities features togehter will get the fusion representation z_t. $Z = [z_1; z_2; \cdots; z_n]$ is the fusion represention of a sentence.

Temporal-Gate We have got the weighted modalities fusion represention by utilizing modality-gate. Then, temporal-gate will compute the weights of every time-steps. Figure 2(b) shows the construction of the temporal-gate. However, there is still a problem that there is no connection between the fusion representation previously obtained. So we use a GRU layer to solve this problem.

$$z_t^{gru} = \text{GRU}(z_t); z_t^{gru} \in R^d \tag{6}$$

z_t^{gru} in the above equation is the hidden state of GRU. We set its length to be d which is the same as the Bi-LSTM layer. The sequence output of the sentence is Z^{gru}=$[z_1^{gru}; z_2^{gru}; \cdots; z_n^{gru}]; Z^{gru} \in R^{n \times d}$.

$$S = \text{softmax}(\mathrm{Z}^{gru\mathrm{T}}); S \in R^{d \times n} \tag{7}$$

$$M = \text{Multiply}(Z^{gru}, S^{\mathrm{T}}); M \in R^{n \times d} \tag{8}$$

We feed Z^{gru} to a softmax layer to calculate the importance of each time-step information which is shown in equation (7). At last, we multiply S with Z^{gru} to get the weighted feature M.

3.3 Prediction

In order to ensure that no information is missed, we use two parts to predict: 1)The output of the temporal-gate. 2)The output of modality-gate. We utilize a max-pooling layer to filter and reduce the temporal-gate features. We use D^{mg} and D^{tg} to represent the outputs of modality and temporal gate, respectively:

$$D^{tg} = \text{MaxPooling}(M); D^{tg} \in R^d \tag{9}$$

For the second part, we use a fully connected layer to filter features,

$$D^{mg} = \tanh(W_m \cdot Z + b_m); D^{mg} \in R^{2d} \tag{10}$$

Where W_m and b_m are the parameters of the fully connected layer. We concatenate D^{tg} and D^{mg} for the final prediction:

$$D = D^{tg} \oplus D^{mg} \tag{11}$$

$$p_\theta(i) = \text{softmax}(\text{W}_p \cdot D + b_p) \tag{12}$$

4 EXPERIMENTATION

4.1 Dateset

We benchmark HGMN's understanding of human communication on three tasks: 1) multimodal speaker traits recognition, 2) multimodal sentiment analysis and 3) multimodal emotion recognition. We perform experimentations on five publicly available datasets and compare the performance of HGMN with the performance of competitive approaches on the same datasets.

Trait Recognition: POM(Persuasion Opinion Multimodal) dataset [6] contains movie review videos annotated for the following speaker traits: confidence, passion, dominance, credibility, entertaining, reserved, trusting, relaxed, nervous and humorous. 903 videos were split into 600 for training, 100 for validation and 203 for testing.

Sentiment Analysis: YouTube dataset [4] contains videos from the social media web site YouTube that span a wide range of product reviews and opinion videos. Out of 46 videos, 30 are used for training, 5 for validation and 11 for testing. **MOUD** To show that HGMN is generalizable to other languages, we perform experimentation on the MOUD dataset [8] which consists of product review videos in Spanish. Each video consists of multiple segments labeled to display positive, negative or neutral sentiment. Out of 79 videos in the dataset, 49 are used for training, 10 for validation and 20 for testing. **ICT-MMMO** dataset consists of online social review videos that encompass a strong diversity in how people express opinions, annotated at the video level for sentiment. The dataset contains 340 multimodal review videos, of which 220 are used for training, 40 for validation and 80 for testing.

Emotion Analysis: CMU-MOSEI [4] is a collection of 22634 opinion video clips. Each opinion video is annotated with sentiment in the range [-3,3]. There are 16188 segments in the train set, 1832 in the validation set and 4614 in the test set.

4.2 Modality Features

Text Modality: All the datasets provide manual transcriptions. We use glove [7] to convert the transcripts of videos into a sequence of word vectors. The dimension of the word vectors is 300. **Vision Modality:** Facet is used to extract a set of features including per-frame basic, advanced emotions and facial action units as indicators of facial muscle movement. **Audio Modality:** We use COVAREP [3] to extract low level acoustic features including 12 Melfrequency

cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients.

Modality Alignment: To reach the same time alignment between different modalities we choose the granularity of the input to be at the level of words. The words are aligned with audio using P2FA [13] to get their exact utterance times. Time-step t represents the t-th spoken word in the transcript. We treat speech pause as a word with vector values of all zero across dimensions. The visual and acoustic modalities follow the same granularity. We use expected feature values across the entire word for vision and acoustic since they are extracted at a higher frequency (30 Hz for vision and 100 Hz for acoustic).

Table 1. Results for sentiment analysis on the ICT-MMMO, YouTube and MOUD dataset

Dataset	ICT-MMMO		YouTube		MOUD	
Task						
Metric	A^2	F1	A^3	F1	A^2	F1
SOTA2	$73.8^{\ddagger}$	$73.1^{\ddagger}$	$51.7^{\ddagger}$	$51.6^{\ddagger}$	$81.1^{\triangle}$	$80.9^{\flat}$
SOTA1	$76.3^{\flat}$	$76.2^{\flat}$	$55.0^{\flat}$	$53.5^{\flat}$	$81.1^{\flat}$	$81.2^{\triangle}$
HGMN	**79.6**	**79.4**	**56.3**	**54.2**	**81.6**	**81.6**
ΔSOTA	$\uparrow$ 3.3	$\uparrow$ 3.2	$\uparrow$ 1.3	$\uparrow$ 0.7	$\uparrow$ 0.5	$\uparrow$ 0.4

4.3 Baseline

SVM (§) a SVM is trained on the concatenated multimodal features for classification or regression [17].
EF-LSTM(♯) concatenates the inputs from different modalities at each time-step and uses that as the input to a single LSTM.
SAL-CNN (○) is a model that attempts to prevent identity-dependent information from being learned by using Gaussian corruption introduced to the neuron outputs.
TFN (•) explicitly models view-specific and cross-view dynamics by creating a multi-dimensional tensor that captures unimodal, bimodal and trimodal interactions across three modalities. It is the current state of the art for CMU-MOSI dataset.
BC-LSTM (†) [9] is a model for context-dependent sentiment analysis and emotion recognition.
DF (♮) [5] is a model that trains one deep model for each modality and performs decision voting on the output of each modality network.
MARN (△) [16] is a model which can discover interactions between modalities through time using a neural component called the Multi-attention Block (MAB)

Table 2. Results for trait recognition on the POM dataset. Human traits use shorthand, for example, Con. represent Confident.

Dataset	POM									
Task	Con.	Pas.	Dom.	Cre.	Ent.	Res.	Tru.	Rel.	Ner.	Hum.
Metric	A^7	A^7	A^7	A^7	A^7	A^5	A^5	A^5	A^5	A^5
SOTA2	$30.0^{\flat}$	$33.0^{\triangle}$	$38.4^{\triangle}$	$31.6^{\flat}$	$33.5^{\triangle}$	$36.9^{\triangle}$	$55.7^{\triangle}$	$52.2^{\triangle}$	$47.3^{\triangle}$	$45.6^{\bullet}$
SOTA1	$34.5^{\ddagger}$	$35.7^{\ddagger}$	$41.9^{\ddagger}$	$34.5^{\ddagger}$	$37.9^{\ddagger}$	$38.4^{\ddagger}$	$57.1^{\ddagger}$	$53.2^{\ddagger}$	$47.8^{\ddagger}$	$47.3^{\ddagger}$
HGMN	**36.4**	**35.9**	**43.9**	**34.7**	**38.7**	**39.4**	**57.6**	**55.7**	**49.8**	**47.8**
ΔSOTA	↑ 1.9	↑ 0.2	↑ 2.0	↑ 0.2	↑ 0.8	↑ 1.0	↑ 0.5	↑ 2.5	↑ 2.0	↑ 0.5
	MAE									
SOTA2	$1.016^{\dagger}$	$0.993^{\ddagger}$	$0.589^{\dagger}$	$0.942^{\dagger}$	$0.927^{\dagger}$	0.879°	0.533°	$0.597^{\natural}$	0.697°	$0.767^{\dagger}$
SOTA1	$0.952^{\ddagger}$	$0.983^{\dagger}$	$0.835^{\ddagger}$	$0.903^{\ddagger}$	$0.913^{\ddagger}$	$0.821^{\ddagger}$	$0.521^{\ddagger}$	$0.566^{\ddagger}$	$0.654^{\ddagger}$	$0.727^{\ddagger}$
HGMN	**0.947**	**0.978**	**0.831**	**0.901**	**0.906**	**0.813**	**0.517**	**0.565**	**0.650**	**0.721**
ΔSOTA	↑ 0.005	↑ 0.005	↑ 0.004	↑ 0.002	↑ 0.007	↑ 0.008	↑ 0.004	↑ 0.001	↑ 0.004	↑ 0.006
	r									
SOTA2	$0.395^{\ddagger}$	$0.428^{\ddagger}$	$0.313^{\ddagger}$	$0.367^{\ddagger}$	$0.395^{\ddagger}$	$0.333^{\ddagger}$	$0.212^{\flat}$	$0.255^{\ddagger}$	$0.318^{\ddagger}$	$0.386^{\ddagger}$
SOTA1	$0.431^{\flat}$	$\mathbf{0.450}^{\flat}$	$0.411^{\flat}$	$0.380^{\flat}$	$0.452^{\flat}$	$0.368^{\flat}$	$0.296^{\ddagger}$	$\mathbf{0.309}^{\flat}$	$0.333^{\flat}$	$0.408^{\flat}$
HGMN	**0.433**	0.444	**0.426**	**0.387**	**0.462**	**0.389**	**0.302**	**0.309**	**0.337**	**0.419**
ΔSOTA	↑ 0.002	↓ 0.006	↑ 0.015	↑ 0.007	↑ 0.010	↑ 0.021	↑ 0.006	-	↑ 0.004	↑ 0.011

and storing them in the hybrid memory of a recurrent component called the Long-short Term Hybrid Memory (LSTHM).

MFN(‡) [15] is a modal that explicitly accounts for both interactions in a neural architecture and continuously models them through time.

GMFN(♭) [18]is a novel multimodal fusion technique called the Graph Memory Fusion Network that dynamically fuses modalities in a hierarchical manner.

4.4 Results and Discussion

Table 1, 2, 3 summarizes the comparison between HGMN and proposed baselines for multimodal traits recognition, sentiment analysis and emotion analysis.

The results of our experiments can be summarized as follows: HGMN achieves the best performance for multimodal human communication comprehension. Table 1 shows the sentiment analysis experiment results of HGMN and other baselines on the three datasets. Our approach has achieved the highest accuracies and F1 in all cases. Table 2 shows the performance of the HGMN on POM dataset, where it achieves the best performance on all 10 speaker trait classification tasks. Table 3 shows the emotion analysis experiment results on the CMU-MOSEI dataset. Our approach has achieved the best WA and F1 in most cases.

Table 3. Results for emotion analysis on the CMU-MOSEI dataset

Dataset	CMU-MOSEI Emotion											
Task	Anger		Disgust		Fear		Happy		Sad		Surpris	
Metric	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1	WA	F1
SOTA2	$60.5^{\bullet}$	$72.0^{\triangle}$	$67.0^{\natural}$	$73.2^{\triangle}$	$60.0^{\triangle}$	**89.9**$^{\triangle}$	$66.3^{\flat}$	$66.6^{\bullet}$	$59.2^{\sharp}$	$61.8^{\triangle}$	$53.3^{\ddagger}$	$85.4^{\ddagger}$
SOTA1	$62.6^{\flat}$	$72.8^{\flat}$	$69.1^{\flat}$	$76.6^{\flat}$	$64.2^{\S}$	**89.9**$^{\flat}$	$66.5^{\bullet}$	$71.0^{\triangle}$	$60.4^{\flat}$	$66.9^{\flat}$	$53.7^{\flat}$	$85.5^{\flat}$
HGMN	**63.1**	**73.0**	**69.9**	**77.4**	**64.6**	**89.9**	**67.2**	**71.5**	**61.1**	**67.2**	**54.4**	**86.5**
ΔSOTA	↑0.5	↑0.2	↑0.8	↑0.8	↑0.4	-	↑0.7	↑0.5	↑0.7	↑0.3	↑0.7	↑1.0

5 Conclusion

This paper introduced a novel approach for multi-modality sequential learning called Hierarchical-gate Multimodal Network(HGMN). Each modality is first encoded by Bi-LSTM which aims to capture the intra-modal interactions within single modality. Subsequently, we merge the independent information of multi-modality using two gate layers. The first gate which is named as modality-gate will calculate the weight of each modality. And the other gate called temporal-gate will capture the importances of every time-step. We also use a GRU layer to capture cross-modal interactions between two gates. We systematically investigate three typical multimodal tasks, i.e., multimodal sentiment analysis, multimodal emotion recognition and speaker traits recognition, to justify the effectiveness of our proposed approach. Detailed evaluation on multiple multimodal benchmark datasets, such as CMU-MOSEI and POM, shows that our proposed approach significantly improves the state-of-the-art.

Acknowledgments

The research work is partially supported by the Key Project of NSFC No.61702149 and two NSFC grants No.61672366, No.61673290.

References

1. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. The annals of mathematical statistics **37**(6), 1554–1563 (1966)
2. Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. pp. 163–171. ACM (2017)
3. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: Covarepa collaborative voice analysis repository for speech technologies. In: 2014 ieee international conference on acoustics, speech and signal processing (icassp). pp. 960–964. IEEE (2014)

4. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 169–176. ACM (2011)
5. Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., Morency, L.P.: Deep multimodal fusion for persuasiveness prediction. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. pp. 284–288. ACM (2016)
6. Park, S., Shim, H.S., Chatterjee, M., Sagae, K., Morency, L.P.: Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In: Proceedings of the 16th International Conference on Multimodal Interaction. pp. 50–57. ACM (2014)
7. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
8. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2539–2544 (2015)
9. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 873–883 (2017)
10. Quattoni, A., Wang, S., Morency, L.P., Collins, M., Darrell, T.: Hidden conditional random fields. IEEE Transactions on Pattern Analysis & Machine Intelligence (10), 1848–1852 (2007)
11. Rajagopalan, S.S., Morency, L.P., Baltrusaitis, T., Goecke, R.: Extending long short-term memory for multi-view structured learning. In: European Conference on Computer Vision. pp. 338–353. Springer (2016)
12. Wörtwein, T., Scherer, S.: What really mattersan information gain analysis of questions and reactions in automated ptsd screenings. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 15–20. IEEE (2017)
13. Yuan, J., Liberman, M.: Speaker identification on the scotus corpus. Journal of the Acoustical Society of America **123**(5), 3878 (2008)
14. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 (2017)
15. Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L.P.: Memory fusion network for multi-view sequential learning. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
16. Zadeh, A., Liang, P.P., Poria, S., Vij, P., Cambria, E., Morency, L.P.: Multi-attention recurrent network for human communication comprehension. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
17. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems **31**(6), 82–88 (2016)
18. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246 (2018)