INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multimodal Sentiment Analysis with Temporal Modality Attention

Fan Qian, Jiqing Han

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{qianfan, jqhan}@hit.edu.cn

Abstract

Multimodal sentiment analysis is an important research that involves integrating information from multiple modalities to identify a speaker underlying attitude. The core challenge is to model cross-modal interactions which span across both the different modalities and time. Although great progress has been made, the existing methods are still not sufficient for modeling cross-modal interactions. Inspired by previous research in cognitive neuroscience that humans perceive intentions through focusing on different modalities over time, in this paper we propose a novel attention mechanism called Temporal Modality Attention (TMA) to simulate this process. Cross-modal interactions are modeled using this human-like TMA mechanism which focuses on specific modalities dynamically as recurrent modeling proceed. To verify the effectiveness of TMA, we conduct comprehensive experiments on multiple benchmark datasets for multimodal sentiment analysis. The results show a consistently significant improvement compared to the baseline models.

Index Terms: temporal modality attention, cross-modal interactions, multimodal sentiment analysis

1. Introduction

Sentiment analysis aims to automatically uncover the underlying attitude that a speaker hold towards an entity, and there is a long history for the research of text-only sentiment analysis. Although this type of sentiment analysis achieves superior performance, it is insufficient for inferring sentiment content from spoken language through words, phrases and their compositionality [1]. By exploring the consistency and complementarity of different modalities, e.g., language, audio and visual modalities, the characterization of sentiment may be more comprehensive and accurate. As a result, there has been a recent tendency which integrates information from multiple modalities to infer factual sentiment, i.e., Multimodal Sentiment Analysis (MSA) [2, 3, 4, 5].

There are two major challenges in MSA. The first one is to model modality-specific interactions. Based on temporal characteristic of spoken language, Recurrent Neural Network (RNN) or other temporal models are used as unimodal encoders to extract more high-level sentiment representations. Furthermore, the second challenge is to model cross-modal interactions which can change the perception of the expressed sentiment in an inconclusive and complicated way [3]. Great contributions have been made by many previous studies relating to modeling cross-modal interactions. They either using tensor products [3] as well as their low-rank variants [6] or associate a relevance score to the memory dimensions of recurrence representations to identify the cross-modal interactions [4, 7]. Recently, multimodal transformer [8] that takes inherent data non-alignment into account and models long-range dependencies between elements across modalities is proved to be efficient. All of these works have a significant push towards more promising multimodal sentiment analysis. However, how to effectively model cross-modal interactions remains an open research problem.

In addition, previous research in cognitive neuroscience have demonstrated the existence of selective attention mechanisms [9, 10], which is a top–down control mechanism that allows us to focus on a particular feature or modality at each moment. In other words, this neural mechanism reveals that humans perceive intentions through focusing on different modalities over time, and the whole intentions are judged by integrating temporal attended information.

Inspired by these studies, we hypothesize that the computational modeling of cross-modal interactions also requires the same neural mechanism. To this end, we propose Temporal Modality Attention (TMA), an attention mechanism for cross-modal interactions which considers the different attention weights required for different modality at each moment. Specifically, at each recurrence time-step, the representations which build upon unimodal encoders are captured, and each modality has one such representation. Given these encoded representations, we argue that there exists a mapping function which can learn the modal weights adaptively. We construct an attention network as the mapping function fed into encoded representations to output modal weights. Meanwhile, for comparison and integration of multimodal sentiment, the encoded representations are mapped to a common sentiment subspace (CSS). And final cross-modal interactions representations are captured by the weighted sum in CSS. Overall, TMA and underlying unimodal encoders jointly model cross-modal and modality-specific interactions.

In order to verify effectiveness of TMA, we conduct comprehensive experiments on three benchmark datasets: CMU-MOSI, ICT-MMMO, and YouTube for multimodal sentiment analysis. The results show a consistently significant improvement compared to the baseline models. We also visualize TMA to display that the learned attention weights at each recurrence time-step have a plausible sentiment preference. Our codes are publicly available at `https://github.com/qianfan1996/TMA`.

2. Related Work

Modeling cross-modal interactions has been a core challenge in multimodal sentiment analysis. Many methods have been reported in the literature to address this issue. Earlier work utilized fusion approaches such as concatenation of input features [11] to obtain multimodal representations. Tensor Fusion Network [3] and its approximate low-rank model [6] are proposed by using Cartesian-products to model cross-modal interactions. Multimodal Factorization Model (MFM) [12] factorizes multimodal representations into multimodal discriminative factors and modality-specific generative factors to deal with the presence of noisy modalities. The models [13] which takes the relation and dependencies among the utterances into account are

http://dx.doi.org/10.21437/Interspeech.2021-487

presented to learn contextual features. Other approaches have been applied by using either attention memory mechanisms [4, 7] or multistage fusion strategies [5] to learn multimodal representations. Recently, transformer-based models [8, 14] have been exploited to capture more representative sentiment information and achieve new state-of-the-art (SOTA) performance.

All these methods assume that cross-modal interactions should be discovered with identical modality contributions. In contrast, our proposed TMA models cross-modal interactions considering different modality contributions. We emphasize that our method has two advantages over previous methods: probably corresponds to the process of humans perception; higher efficiency and better performance.

3. Methodology

To describe proposed TMA comprehensively, we choose Multi-attention Recurrent Network (MARN) [4] as the backbone model and replace internal Multi-attention Block (MAB) module with our TMA for cross-modal interactions. The integrated model framework is shown in Figure 1.

3.1. Long-short Term Hybrid Memory

We briefly describe the modality-specific interactions module called Long-short Term Hybrid Memory (LSTHM) within MARN. LSTHM is quite similar to Long-short Term Memory (LSTM) [15], except for adding a memory factor which carries cross-modal interactions information. Specifically, given a set of three modalities $M = \{l(anguage), a(udio), v(isual)\}$ in the domain of the data, subsequently three LSTHMs are built in the pipeline. For each modality $m \in M$, the input to the m modality LSTHM is the form $\boldsymbol{X}^m = \{\boldsymbol{x}_1^m, \boldsymbol{x}_2^m, ..., \boldsymbol{x}_T^m; \boldsymbol{x}_t^m \in \mathbb{R}^{d_x^m}\}$, where d_x^m is the dimensionality of the input $\boldsymbol{x}_t^m$ of modality m. The Long-short Term Hybrid Memory is formulated as follows,

$$
\begin{aligned}
\boldsymbol{f}_t^m &= \sigma(W_f^m \boldsymbol{x}_t^m + U_f^m \boldsymbol{h}_{t-1}^m + V_f^m \boldsymbol{z}_{t-1} + \boldsymbol{b}_f^m) \\
\boldsymbol{i}_t^m &= \sigma(W_i^m \boldsymbol{x}_t^m + U_i^m \boldsymbol{h}_{t-1}^m + V_i^m \boldsymbol{z}_{t-1} + \boldsymbol{b}_i^m) \\
\boldsymbol{o}_t^m &= \sigma(W_o^m \boldsymbol{x}_t^m + U_o^m \boldsymbol{h}_{t-1}^m + V_o^m \boldsymbol{z}_{t-1} + \boldsymbol{b}_o^m) \\
\hat{\boldsymbol{c}}_t^m &= tanh(W_c^m \boldsymbol{x}_t^m + U_c^m \boldsymbol{h}_{t-1}^m + V_c^m \boldsymbol{z}_{t-1} + \boldsymbol{b}_c^m) \\
\boldsymbol{c}_t^m &= \boldsymbol{f}_t^m \odot \boldsymbol{c}_{t-1}^m + \boldsymbol{i}_t^m \odot \hat{\boldsymbol{c}}_t^m \\
\boldsymbol{h}_t^m &= \boldsymbol{o}_t^m \odot tanh(\boldsymbol{c}_t^m)
\end{aligned}
\tag{1}
$$

where $\boldsymbol{h}_{t-1}^m$ is the output of m modality LSTHM at time-step $t-1$, $\boldsymbol{z}_{t-1}$ is the latent code which acts as a hybrid memory factor, allowing individual LSTHM to carry cross-modal interactions information, $\odot$ denotes the Hadamard product (element-wise product), and σ is the sigmoid activation function.

3.2. Temporal Modality Attention

TMA is designed for cross-modal interactions and is the core of our work. As demonstrated in previous cognitive neuroscience literature, at each time-step, human brains focus on a particular feature or modality. TMA simulates this process, and different modalities are endowed with different attention weights as recurrent modeling proceed.

As illustrated in Figure 1, at each time-step of TMA recursion, the representation $\boldsymbol{h}_t^m$ (the block in the red boxes) which builds upon unimodal encoder LSTHM is mapped to a common sentiment subspace (CSS),

$$
\boldsymbol{s}_t^m = \mathcal{D}_m(\boldsymbol{h}_t^m) \tag{2}
$$

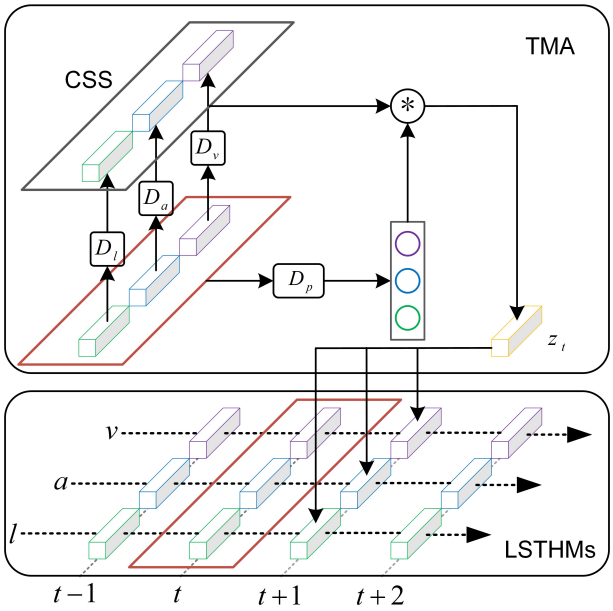

Figure 1: *Overview of the model pipeline*

where $\boldsymbol{s}_t^m$ is the CSS representation of modality m, $\mathcal{D}_m$ is the mapping function which aims to further extract sentiment-related information from the unimodal encoder. Furthermore, the representations in the same sentiment subspace are allowed for comparison and integration of sentiment.

Then, the different modal weights are learned adaptively according to the current encoded information of three modalities,

$$
\boldsymbol{\alpha}_t = \mathcal{D}_p(\bigoplus_{m \in M} \boldsymbol{h}_t^m) \tag{3}
$$

where $\mathcal{D}_p$ denotes the attention network, in which the underlying hypothesis is that it can learn the modal weights given the encoded representations via abundant training data. $\boldsymbol{\alpha}_t$ ($[\alpha_t^l, \alpha_t^a, \alpha_t^v]$) are softmax activated scores for each modality at time-step t. $\bigoplus$ denotes the concatenation of representations. To explore the distribution of sentiment of all modalities accurately, we select encoded representations $\boldsymbol{h}_t^m$ instead of the sentiment subspace representations $\boldsymbol{s}_t^m$ as the input of $\mathcal{D}_p$ preventing lack of information. Applying softmax at the output layer allows for regularizing score vectors over the $\boldsymbol{h}_t^m$.

Final temporal representation is defined as,

$$
\boldsymbol{z}_t = \sum_{m \in M} \alpha_t^m \boldsymbol{s}_t^m \tag{4}
$$

where $\boldsymbol{z}_t$ is the attended cross-modal interactions representation, which carries the sentiment-related information and is applied to modality-specific interactions.

The last time-step outputs of TMA and LSTHMs are concatenated together to pass through a single hidden layer Fully Connected Neural Network (FCNN) for classification of sentiment. In summary, TMA and underlying temporal encoders LSTHMs jointly model cross-modal and modality-specific interactions, and the integrated framework is differentiable end-to-end which allows the network parameters to be optimized using gradient descent approaches.

Table 2: *Results for multimodal sentiment analysis on CMU-MOSI, ICT-MMMO, and YouTube datasets*

Dataset	CMU-MOSI					ICT-MMMO		YouTube	
Metric	Acc(2)	F1	Acc(7)	MAE	r	Acc(2)	F1	Acc(3)	F1
Majority	50.2	50.1	17.5	1.864	0.057	40.0	22.9	42.4	25.2
RF [16]	56.4	56.3	21.3	-	-	70.0	69.8	49.3	49.2
C-MKL [17]	72.3	72.0	30.2	-	-	80.0	72.4	50.2	50.8
EF-LSTM [15]	73.3	73.2	32.4	1.023	0.622	80.0	78.5	44.1	43.6
MV-LSTM [18]	73.9	74.0	33.2	1.019	0.601	72.5	72.3	45.8	43.3
BC-LSTM [13]	73.9	73.9	28.7	1.079	0.581	70.0	70.1	47.5	47.3
TFN [3]	74.6	74.5	28.7	1.040	0.587	72.5	72.6	47.5	41.0
MFN (baseline) [7]	76.1	72.2	33.4	0.975	0.638	80.0	83.7	57.6	56.6
MFN (TMA)	**77.4**	**72.6**	**34.7**	**0.960**	**0.652**	**85.0**	**87.8**	**61.0**	**60.0**
MARN (baseline) [4]	76.4	71.8	**33.6**	0.975	0.645	81.3	85.7	54.2	53.8
MARN (TMA)	**78.1**	**73.6**	33.4	**0.956**	**0.646**	**85.0**	**88.7**	**59.3**	**58.0**

4. Experimental Setup

4.1. Datasets

All datasets consist of monologue videos. The speaker intentions are conveyed through three modalities: language, audio, and visual.

CMU-MOSI dataset [19] is a collection of video opinions from YouTube movie reviews. It consists of 2199 segments, each of which is annotated with a sentiment score from -3 (strongly negative) to +3 (strongly positive). We split the dataset into train (1284), validation (229) and test (686) set.

ICT-MMMO dataset [20] is a collection of videos about movie reviews on social media sites in the form of a person speaking directly to the camera, expressing their comments about the movie. It is annotated with a sentiment score from +1 (strongly negative) to +5 (strongly positive) and split into train (220), validation (40) and test (80) set.

YouTube dataset [2] contains a series of videos on diverse topics such as toothpaste, camera and baby products. Out of 269 videos, 169 are used for training, 41 for validation and 59 for testing.

The train, validation and test split details of all datasets are showed in Table 1. We follow the speaker independent setting to verify the generalization of models.

Table 1: *Data splits to ensure speaker independent learning*

Dataset	CMU-MOSI	ICT-MMMO	YouTube
# Train	1284	220	169
# Valid	229	40	41
# Test	686	80	59

4.2. Features

Human multimodal behaviors are mainly composed of three modalities: language, audio and visual. To get the exact utterance timestamp of each word, we use P2FA toolkit [21] to perform forced alignment, which allows us to align the three modalities together. Since words are the basic units of language, we utilize the time interval of each word as a time-step, and we then calculate the average features on the audio and visual modalities over the word time ranges. We split or pad each video into 20 words and acquire three feature sequences of length 20. The feature extraction process of individual modality is described as follows.

For the language modality, GloVe [22] word embeddings pretrained on 840 billion tokens are extracted to convert the transcriptions of spoken language to word vectors sequence. The dimension of word vectors is 300.

For the audio modality, COVAREP [23] is used to extract low-level acoustic features. The acoustic features include 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. The dimension of audio features is 74.

For the visual modality, the library Facet [24] is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features [25, 26]. The dimension of visual features is 35.

4.3. Comparison Metrics

Different evaluation tasks are performed for different datasets based on the provided labels. For classification, we report F1 score and accuracy Acc(k) where k denotes the number of classes. For regression, we report Mean Absolute Error (MAE) and Pearson's correlation r. Higher values denote better performance for all the metrics, except MAE where lower values denote better performance.

4.4. Baseline Models

To verify effectiveness of proposed TMA, we choose two baseline models and replace their cross-modal interactions modules with TMA.

Memory Fusion Network (MFN) [7] encodes each modality independently for modality-specific interactions using a component called the System of LSTMs, while Delta-memory Attention Network (DMAN) module in MFN is designed for cross-modal interactions by associating a relevance score to the memory dimensions of each LSTM. We replace the DMAN module with our TMA and compare the integration models with MFN.

Multi-attention Recurrent Network (MARN) [4] encodes each modality for modality-specific interactions with proposed Long-short Term Hybrid Memory (LSTHM) which involves integating cross-modal representations into unimodal modeling. And cross-modal interactions are discovered at each recurrence time-step using a specific neural component call Multi-attention Block (MAB). We replace the MAB module with our TMA and compare the integration models with MARN.

In addition, based on a fair comparison, we list some common models using LSTM for sequence modeling instead of

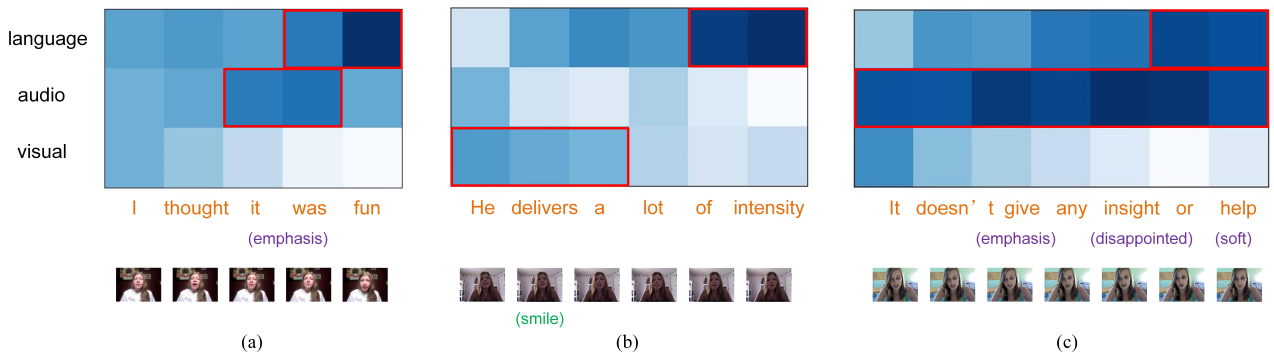

Figure 2: *Visualization of learned modality weights on CMU-MOSI dataset. The darker the blue, the greater the weight. The red boxes emphasize specific moments of interest. The orange font represents spoken text. The purple and green font with parentheses represent intention conveyed by audio and facial expression, respectively.*

transformer-based models, such as Early Fusion LSTM (EF-LSTM) [15], Multi-view LSTM (MV-LSTM) [18], Bidirectional Contextual LSTM (BC-LSTM) [13] as well as Tensor Fusion Network (TFN) [3].

5. Results and Discussion

5.1. Performance

Table 2 presents the performance of our models compared with other methods and baseline models for multimodal sentiment analysis. MFN (TMA) which replace the cross-modal interactions module within MFN with our TMA consistently improve over MFN baseline in all metrics, and MARN (TMA) outperforms MARN baseline in all metrics except 7-class accuracy. These observations highlight TMA capability in human multimodal computational modeling. We believe that our proposed TMA can easily be enhanced with more advanced unimodal temporal encoders such as transformer [27] and BERT [28].

5.2. Efficiency

Since different datasets have different labels, the same model architectures also have subtle differences. Table 3 shows the number of parameters of models on CMU-MOSI, ICT-MMMO and YouTube datasets. The unit is "K" meaning thousand. We observe that the models with proposed TMA have fewer parameters on all datasets, in particular MARN (TMA) which has almost twice fewer parameters than its counterpart, which further demonstrate the efficiency of TMA.

Table 3: *number of parameters (K) on various datasets*

Dataset	CMU-MOSI	ICT-MMMO	YouTube
MFN	379	501	618
MFN (TMA)	**365**	**472**	**425**
MARN	749	900	945
MARN (TMA)	**377**	**631**	**448**

5.3. Visualization

To understand how TMA works while modeling multimodal data, we choose the same three video clips on CMU-MOSI dataset as [5] and visualize the learned modality weights at each recurrence time-step (see Figure 2).

In Figure 2(a), the language modality is highlighted corresponding to the utterance of the word "fun" that is highly indicative of sentiment at last time-step. Furthermore, we also notice that the person spoke with an emphatic tone dominating sentiment from $t = 3$ to $t = 4$. In Figure 2(b), the model initially focuses on facial expressions because the person is smiling, conveying a positive sentiment. However, at last two time-steps, the model puts more emphasis on language modality as the person utters the word "intensity". In Figure 2(c), the person speak with an emphatic, disappointed and soft tone across all the time-steps in which sentiment preference is conveyed explicitly. Meanwhile, at the last few moments, language modality also plays an important role for implicitly conveying sentiment information.

More importantly, we observe that the variation of attention weights is smooth in three video clips, process of which may correspond to perception of human brain. Furthermore, we find that language is the most informative modality for sentiment analysis that the same observations are obtained in [29, 30]. All these discoveries demonstrate that the learned attention weights can indeed reflect factual sentiment preference, and modeling this process is promising for sentiment analysis as well as natural Human-Computer Interactions (HCI).

6. Conclusions

In this paper, we propose a novel attention mechanism called temporal modality attention (TMA), where different modalities are endowed with different attention weights at each recurrence time-step. Our approach is inspired by previous research in cognitive neuroscience, and we hypothesize that the computational modeling of cross-modal interactions also requires the same neural mechanism. Our experiments verify the effectiveness and efficiency of TMA, and the visualizations of TMA show that the weights learned at each recurrence time-step have a plausible sentiment preference.

7. Acknowledgements

This research is supported by National Key Research and Development Program of China under Grant No.2017YFB1002102 and National Science Foundation of China under Grant No.U1736210.

8. References

[1] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI*, 2019.

[2] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in *ICMI '11*, 2011.

[3] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.

[4] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2018, pp. 5642–5649, 2018.

[5] P. P. Liang, L. Ziyin, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *EMNLP*, 2018.

[6] Z. Liu, Y. Shen, V. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018.

[7] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," *ArXiv*, vol. abs/1802.00927, 2018.

[8] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, pp. 6558–6569, 2019.

[9] E. Macaluso, C. Frith, and J. Driver, "Directing attention to locations and to sensory modalities: multiple levels of selective processing revealed with pet." *Cerebral cortex*, vol. 12 4, pp. 357–68, 2002.

[10] J. L. Mozolic, C. Hugenschmidt, A. Peiffer, and P. Laurienti, "Modality-specific selective attention attenuates multisensory integration," *Experimental Brain Research*, vol. 184, pp. 39–52, 2007.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *ICML*, 2011.

[12] Y.-H. H. Tsai, P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *ArXiv*, vol. abs/1806.06176, 2019.

[13] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *ACL*, 2017.

[14] Z. Wang, Z. Wan, and X. jun Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," *Proceedings of The Web Conference 2020*, 2020.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2004.

[17] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *EMNLP*, 2015.

[18] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *ECCV*, 2016.

[19] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *ArXiv*, vol. abs/1606.06259, 2016.

[20] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, pp. 46–53, 2013.

[21] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, pp. 3878–3878, 2008.

[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.

[23] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep — a collaborative voice analysis repository for speech technologies," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, 2014.

[24] iMotions, "Facial expression analysis," 2017.

[25] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience." *Journal of Personality and Social Psychology*, vol. 39, pp. 1125–1134, 1980.

[26] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, pp. 169–200, 1992.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[29] M. Chen, S. Wang, P. P. Liang, T. Baltrusaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017.

[30] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl. Based Syst.*, vol. 161, pp. 124–133, 2018.