Information Fusion 92 (2023) 37–45

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis

Kyeonghun Kim [1], Sanghyun Park [*,1]

Department of Computer Science, Yonsei University, Korea

ARTICLE INFO

Keywords:
Multimodal Sentiment Analysis
Single-stream Transformer
Multimodal Masked Language Model
Alignment Prediction

ABSTRACT

Multimodal sentiment analysis utilizes various modalities such as Text, Vision and Speech to predict sentiment. As these modalities have unique characteristics, methods have been developed for fusing features. However, the overall modality characteristics are not guaranteed, because traditional fusion methods have some loss of intra-modality and inter-modality. To solve this problem, we introduce a single-stream transformer, All-modalities-in-One BERT (AOBERT). The model is pre-trained on two tasks simultaneously: Multimodal Masked Language Modeling (MMLM) and Alignment Prediction (AP). The dependency and relationship between modalities can be determined using two pre-training tasks. AOBERT achieved state-of-the-art results on the CMU-MOSI, CMU-MOSEI, and UR-FUNNY datasets. Furthermore, ablation studies that validated combinations of modalities, effects of MMLM and AP and fusion methods confirmed the effectiveness of the proposed model.

1. Introduction

Sentiment analysis [1] is the task of extracting sentiments or opinions of people from text. Companies identify which messages or conversations that trigger customer emotions via sentiment analysis to offer better services. Recently, social media platforms, such as Facebook, video lectures, and YouTube have been used to express personal opinions, resulting in a significant increase in multimodal data [2, 3, 4]. Multimodal data refer to the complex information provided along with Vision (facial expressions), Speech (vocal expression or tone), and Text. Each modality that represents the same segment complements other modalities and provides additional information. For example, sentiment in objective sentences such as "The package is red." is difficult to determine. However, in conjunction with facial expressions or voice tone information, sentence sentiments can be predicted more easily. Therefore, Multimodal Sentiment Analysis (MSA) is more accurate than sentiment analysis using only Text.

Multimodal research aims to minimize the loss of information by fusing different modalities. In a previous study, Early-Fusion LSTM (EF-LSTM) and Late-Fusion LSTM (LF-LSTM) were used in the input phase and prediction phase to fuse information from different modalities. However, these methods are ineffective for reflecting inter-modality and intra-modality [2]. In addition, MCTN [5] generated a joint representation with Recurrent Neural Network (RNN) model, which led to a long-term dependency problem. The transformer models based on the attention mechanism [6] have been proposed recently in NLP research to mitigate the long-term dependency problem. In particular, the cross-transformer method can minimize the loss of information and solve the long-term dependency problem [3, 4, 7]. However, because the cross-transformer method assumes the equal contribution of each modality, it uses less textual information than other methods; thus satisfactory performance is difficult to achieve [8]. The gated mechanism was used to adjust the ratio of modality contributions, requiring more computation than previous methods [3]. Therefore, advanced methods are needed to efficiently combine the information from multi-modalities.

A single-stream transformer was proposed to solve the cross-transformer problems in other multimodal research, e.g., data2vec [9], speech T5 [10], and multimodal BERT-based dialogue systems [11]. In these studies, models were trained to learn the two modalities using a single transformer. The most important modalities are not always the same, even if a statement appears in multiple forms. Therefore, it is useful to use as much multi-modal information as possible for sentiment analysis. However, it is difficult to simply increase the number of modalities. To address this problem, we propose AOBERT (All-modalities-in-One BERT), which is a single-stream transformer that can handle three modalities (Text, Vision, and Speech) as inputs to one network for sentiment analysis and emotion detection. We also introduce methods for generating joint representations that can reflect the characteristics of each modality by using Multimodal Masked Language Modeling

* Corresponding author.
E-mail address: sanghyun@yonsei.ac.kr (S. Park).
[1] Source code: https://github.com/kimkyeonghun/MSA

https://doi.org/10.1016/j.inffus.2022.11.022
Received 26 January 2022; Received in revised form 17 November 2022; Accepted 19 November 2022
Available online 22 November 2022

(MMLM) and Alignment Prediction (AP) which are inspired by BERT [12].

The main contributions of AOBERT are as follows: 1) We propose a single-stream transformer model that learns three modalities over a single network. 2) The proposed model utilizes MMLM and AP tasks inspired by BERT to learn joint representations. 3) Experimental results indicate that the proposed model outperforms the state-of-the-art models on three benchmark datasets for sentiment analysis and emotion detection.

2. Related works

2.1. Sentiment analysis

Sentiment analysis [1] is the task of extracting sentiments or opinions of people from text. Traditional NLP methods, such as bag-of-words, have disadvantages with long sentences or specialized text. By incorporating recurrent deep learning methods and fine-tuning large-sized language models, such as BERT and GPT-3 [13], these problems can be overcome. Recently, Aspect-Based Sentiment Analysis (ABSA) [14] and domain-specific sentiment analysis [15, 16] have been studied. ABSA is a complex task to identify sentiments associated with specific aspects of text. Domain-specific sentiment analysis is usually applied in the finance and bio-medical fields.

2.2. Multimodal sentiment analysis

MSA primarily focuses on integrating multiple resources, such as textual, visual, and speech information, to comprehend human emotions. Combining features is necessary because they offer parallel information for the same source and are useful for the disambiguation of affective behavior. Previous studies on MSA have usually focused on the factors of variation that have a more direct correlation with emotion. For example, Sentic Blending [17] fuses the modalities to grasp the emotion associated with multimodal content.

Multimodal research was performed by fusing the data at the input and prediction stage such as Early-Fusion [18] and Late-Fusion [19]. Early-Fusion integrates the functions of each modality in the input stage. However, it can suppress interactions within a modality and cause the modalities have synchronization problems that must exist at the same time steps. Late-Fusion comprises each modality as an independent model and applies a majority-voting or weighted average approach to the model's results. Because each modality exists as an independent network, the interactions between modalities are ignored.

Studies have been conducted to obtain a joint representation to solve these problems. Graph-MFN [20] uses a graph architecture and methodology for memory usage to obtain a joint representation. MCTN obtains a joint representation using hierarchical cyclic translation. However, because the aforementioned models use the RNN architecture, there is loss of information due to the long-term dependency problem. In recent years, the NLP domain has used a transformer model based on self-attention to address the long-term dependency problem. For adopting in multimodal domain, transformer is used as cross-transformer. As a result, the performance is further improved by employing the cross-transformer. For instance, MTMM-ES [21] predicts sentiment and emotion simultaneously in a multi-task learning framework using contextual inter-modal attention.

The single-stream transformer learned about all-modalities-in-one-network to address the problems of cross-transformer in other multimodal research. For example, data2vec [9] uses self-supervised learning based on the teacher forcing and masking method. It exhibited results in various multimodal domains. SpeechT5 [10], which uses speech and text as one input, performs well in numerous speech domains. The multi-modal BERT-based dialogue system [11] generates the multi-modal response according to the context.

3. Methods

3.1. Problem definition

In this study, sentiment analysis and emotion detection were conducted using three modalities: Text, Vision, and Speech. The input modality X is defined as follows:

$$X_T \in R^{d_T \times L},\ X_V \in R^{d_V \times L},\ X_S \in R^{d_S \times L}$$

where X_T, X_V, and X_S refer to Text, Vision, and Speech, respectively. These are vectors of length L with dimensions of d_T, d_V and d_S, respectively. Because L is the fixed length of the input size, certain inputs that are smaller than L would be contained zero padding to fit the size. AOBERT only uses pairs of modalities such as (X_T, X_V), (X_T, X_S) where Text is employed as the anchor modality. The pairs of modalities are defined as follows:

$$T = (X_T)\ V^{'} = (X_T,\ X_V)\ S^{'} = \ (X_T,\ X_S)$$

where T, $V^{'}$, and $S^{'}$ are processed simultaneously in one training step. Next, pairs of modalities undergo backpropagation simultneously. The outputs of model consist of two results: sentiment and emotion. The sentiment (Y_S) and emotion (Y_E) are based on Text labels. Y_S is a real number in the range of [-3, +3] and Y_E is classified as either 1 or 0, which indicates whether to appear or not. Details regarding the outputs of sentiment and emotions are presented in Section 4.1.

3.2. The overall architecture

The overall architecture of AOBERT is shown in Fig. 1. The proposed model can be divided into three parts. The first part is "Joint Embedding", which concatenates pairs of modalities such as $V^{'}$ and $S^{'}$ before AOBERT part. $V^{'}$ and $S^{'}$ are generated by "Fusion Gate" in Joint Embedding. The second part is AOBERT which uses a single-stream transformer model for Text, Vision, and Speech modalities. Within AOBERT, the model is pre-trained on two tasks simultaneously: MMLM and AP. MMLM is inspired by Masked Language Model (MLM) in vanilla BERT and can handle multimodal data. AP is similar to the Next Sentence Prediction (NSP) in vanilla BERT. NSP is a task that understands sentence relationships. Similarly, AP can understand modality relationships by predicting whether multimodal data are paired or not. The result of AOBERT is a joint representation that reflects the characteristics of modalities using MMLM and AP. The final part is a classifier

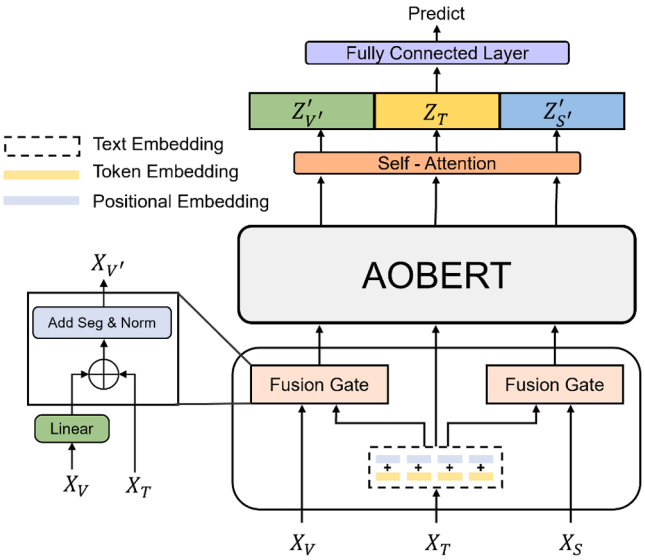

Fig. 1. Overall architecture of AOBERT.

for sentiment analysis and emotion detection.

3.2.1. Joint embedding

Joint Embedding consists of text embedding and Fusion Gate. Text embedding consists of token embedding and positional embedding which convert word tokens of Text X_T to real numbers to provide positional information to Text, respectively. However, because Vision X_V and Speech X_S do not have sequential characteristics, t positional embedding is not required. The proposed model uses the vanilla BERT token embedding and positional embedding and the output of text embedding is used as the input of Fusion Gate.

3.2.2. The fusion gate

AOBERT uses three pairs T, $V^{'}$ and $S^{'}$ as input. $V^{'}$ and $S^{'}$ are generated by Fusion Gate and T is used as the anchor modality. First, a linear layer matches the dimensions between Text and the other modalities. Subsequently, two different modalities are concatenated, and segment embedding is added to distinguish them. Finally, layer normalization is performed. The structure is as follows:

$$A \oplus B = ([A; B^{'}] + Segment) \tag{1}$$

$$B^{'} = Linear(B) \tag{2}$$

$$X_{V^{'}} = LN(X_T \oplus X_V),\ X_{S^{'}} = LN(X_T \oplus X_S) \tag{3}$$

where $A \oplus B$ is defined as "Fuse A and B", where A represents text modality, and B represents another modality, such as Vision or Speech, in (1). The modalities have different dimensions owing to their characteristics. However, because AOBERT sets Text modality as the anchor modality, the dimensions of the other modalities change according to Text. Specifically, B is projected through *Linear* in (2) to concatenate with A. After A and $B^{'}$ are combined according to the length of the sequence in the term $[A; B^{'}]$, segment embedding (*Segment*) is added. Segment embedding is used to distinguish the sentence to which the tokens belong to vanilla BERT. Therefore, it consists of 0 and 1 for A and B. Similarly, AOBERT uses Segment Embedding to distinguish the modality. Finally, *LN* is the LayerNorm with regularization dimension d_T. $X_{V^{'}}$ and $X_{S^{'}}$ have dimensions and lengths of $R^{T\times\ 2L}$. By contrast, X_T has length L.

3.3. AOBERT

The internal structure of AOBERT is shown in Fig. 2. The E tokens are the result of Joint Embedding, and the output of AOBERT is H tokens. Among the H tokens, $H_{\langle CLS\rangle}$ utilizes at Pooler to use downstream tasks and Pooler uses a fully connected layer and tanh activation function. After the three pairs, T, $V^{'}$, and $S^{'}$ are processed by Pooler, we can obtain Z_T, $Z_{V^{'}}$ and $Z_{S^{'}}$ as follows:

$$Z_{\{T,\ V^{'},\ S^{'}\}} = \tanh(Linear\left(H_{<CLS>_{\{T,\ V^{'},S^{'}\}}}\right) \tag{4}$$

3.3.1. Multimodal masked language modeling (MMLM)

The MMLM task is similar to the MLM task in BERT. However, the main difference lies in capturing the dependency between Text and another modality by masking only Text. During pre-training, the input tokens are randomly masked with a 15% probability, and then the masked tokens are replaced with a special token as [MASK], a random token, or kept unchanged with probabilities of 80%, 10%, and 10%, respectively. Finally, the model is trained to predict masked tokens based on the unmasked text tokens and other tokens from other modalities.

3.3.2. Alignment prediction (AP)

In addition to MMLM, we proposed the AP task which was inspired by the NSP task in BERT. Because $V^{'}$ and $S^{'}$ contain two different modalities, AP is applied, to apprehend the relationship between the modalities. For example, X_T and X_V in pair $V^{'}$ are selected from the training data. X_V is a pair with the actual X_T with a probability of 50% (IsPair), whereas X_V is not a pair with the actual X_T with a probability of 50% (UnPair). "IsPair" and "UnPair" denote 1 and 0, respectively.

3.4. Classifier and final prediction

To obtain a meaningful joint presentation, a self-attention layer applies on Z_T, Z_V and Z_S, which is derived from $H_{\langle CLS\rangle}$ for each set of pairs. The [CLS] token is used only for classification tasks. The self-attention layer is illustrated in Fig. 3, which is an example of Z_T. Finally, the multimodal joint representation for each pair is concatenated and fed to a fully connected layer to predict the results. The following formula is used:

$$Z_T^{'} = Linear(\mathrm{ReLU}(Linear(Z_T \oplus Z_T))) \cdot Z_T \tag{5}$$

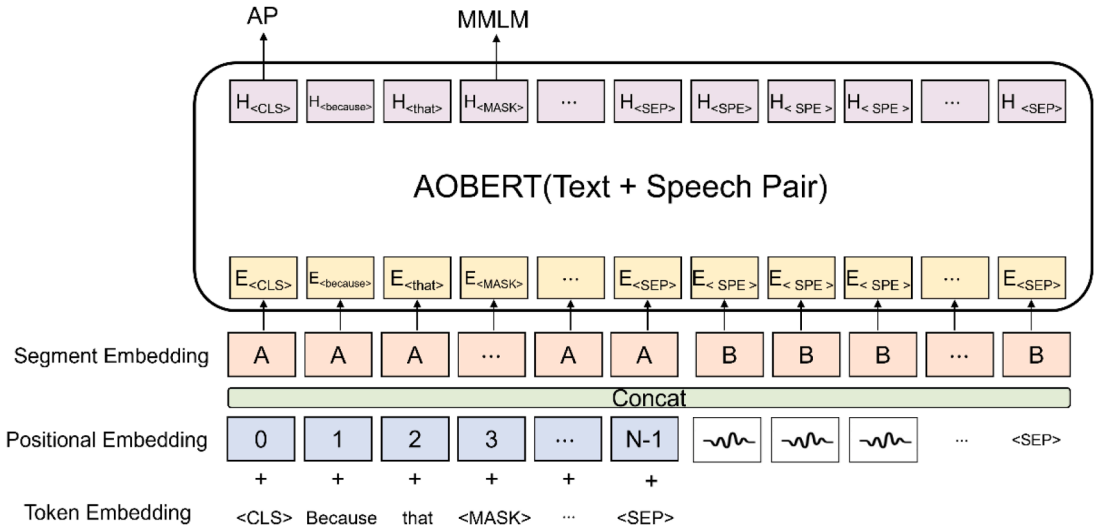

Fig. 2. The Internal Structure of AOBERT when Text and Speech are concatenated.

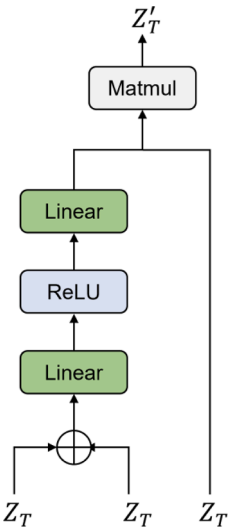

Fig. 3. The Self-attention Layer in Classifier.

3.5. Loss Function

Sentiment analysis and emotion detection use different loss functions, which consist of the sum of L_{joint} and L_{task}. L_{joint} refers to the joint loss function and comprises MMLM and AP pre-training tasks. The MMLM and AP tasks use CrossEntropy loss function and are denoted as L_{MMLM} and L_{AP}, respectively. L_{joint} is calculated as follows:

$$L_{joint} = L_{MMLM} + L_{AP} \tag{6}$$

Because L_{joint} is calculated in three pairs, it has three types: L_T, $L_{V'}$, and $L_{S'}$. Additionally, L_T does not have L_{AP}. The final loss is calculated as the sum of L_{task} and the average of the three joint losses. Overall learning of the model is performed by minimizing (7).

$$L = L_{task} + \frac{(L_T + L_{V'} + L_{S'})}{3} \tag{7}$$

L_{task} is a loss that depends on the task. As sentiment analysis is a regression task, L_{task} uses the Mean Squared Error (MSE) loss function. By contrast, CrossEntropy loss function is utilized for emotion and humor detection.

$$L_{task} = -\frac{1}{N}\sum_{i=0}^{N-1} y_i \cdot log\hat{y}_i = \frac{1}{N}\sum_{i=0}^{N-1} \| y_i - \hat{y}_i \|_2^2 \tag{8}$$

4. Experiment

4.1. Dataset

In this study, the proposed AOBERT was evaluated on the CMU-MOSI [22], CMU-MOSEI [20], and UR-FUNNY [23] datasets for MSA and emotion detection. The information on the dataset is in Table 1.

4.1.1. CMU-MOSI

CMU-MOSI is a widely used dataset in MSA research. It is a collection of YouTube monologues, where speakers express their opinions on movie topics. Sentiment is annotated with a continuous opinion score between [− 3, 3] using the Stanford Sentiment Treebank annotation method [24], and the sentiment − 3/+3 indicates a strong *negative/-positive*. The dataset was divided into training, validation, and test sets; the corresponding number of utterances were 1284, 229, and 686 respectively.

4.1.2. CMU-MOSEI

CMU-MOSEI is an extension of the CMU-MOSI dataset with more utterances and a wider variety of samples, speakers, and topics. It has 23,453 annotated sentences from 1,000 distinct speakers on 250 different issues. Sentiment is annotated using the same method that is employed for CMU-MOSI. The emotions consist of Happy, Sad, Angry, Fear, Disgust, and Surprise by Ekman Emotions [25] and the range of annotated scores is [0, 3], where 0 indicates no emotion and 3 indicates high emotion. In addition, emotion is a multi-emotion label that can appear simultaneously in one utterance. CMU-MOSEI was divided into training, validation, and test sets, the corresponding number of utterances were 16216, 1871, and 4653, respectively.

4.1.3. UR-FUNNY

We used the UR-FUNNY dataset, which provides multimodal utterances that act as punchlines sampled from TED Talks, for Multimodal Humor Detection (MHD). TED Talks are the most diverse idea-sharing channel with regard to both speakers and topics. They span a broad spectrum of humor owing to the diversity of speakers and topics. UR-FUNNY contained 1866 videos and transcripts from TED. These videos were selected from 1741 speakers and 417 topics. A laughter markup was used to filter out 8,257 humorous punchlines from the transcripts. Each target utterance was assigned a binary label for *humor/non-humor*. The sizes of the training, validation, and test sets were 7614, 980, and 994, respectively.

4.2. Evaluation Metrics

For both CMU-MOSI and CMU-MOSEI, sentiment analysis is a regression task with Mean Absolute Error (MAE) as a metric. In addition, it has classification scores, such as seven-class accuracy (A7), which ranges from − 3 to 3, binary accuracy (A2), and F1-score. The binary accuracy (A2) has been used in two different ways. Before 2019, the models classified sentiments as *negative* if the sentiment score was < 0, and *non-negative*, otherwise. In recent studies, binary accuracy has been calculated on a more precise measure of *negative/positive* classes where $<$ 0 corresponds to *negative* and > 0 corresponds to *positive*. There is a difference of approximately 2% - 3% between these two measures, with the latter having a higher accuracy. We report both measures using the segmentation marker - / -, where the left side is for *neg./non-neg.* and the right side is for *neg./pos* [26].

For emotion detection in the CMU-MOSEI dataset, as it is a classification task for each emotion, A2 and F1 were used. By contrast, for the UR-FUNNY dataset, the task was a standard binary classification with binary accuracy (A2) as the evaluation metric.

4.3. Feature extraction

We used the CMU-Multi-modal Data SDK [27] for feature extraction about each modality except UR-FUNNY. In UR-FUNNY, the provided features were used.

4.3.1. Text features

In previous studies, GloVe word embedding [28] was used as the text modality feature for each token. However, state-of-the-art results have recently been obtained by utilizing pre-trained BERT as a feature extractor for textual utterances. Therefore, we use (B) to denote the BERT embedding which has dimensions of 1024 for fair comparisons in the tables in Section 5.

4.3.2. Visual features

Both CMU-MOSI and MOSEI use Facet [29] to extract facial action units and record facial muscle movements to represent basic and advanced emotions in each frame. UR-FUNNY extracts the features of facial expressions using OpenFace2 [30], a facial behavior analysis toolkit. The visual feature dimensions were 47, 35, and 75 for CMU-MOSI, MOSEI, and UR-FUNNY, respectively.

4.3.3. Speech features

All datasets utilize COVAREP [31], a speech processing algorithm for extracting low-level speech features. The features include 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking, voiced/unvoiced segmenting features, glottal source parameters, and peak slope parameters. The dimension of speech features was 74.

4.4. Experiment setup

We trained AOBERT using a *bert-large-uncased* pre-trained model. Specifically, we used the Adam optimizer with a learning rate of 5e-5 or 5e-4 and trained the network for 100 epochs with EarlyStopping. We set *max_seq_len* to 40 for CMU-MOSI and 50 for CMU-MOSEI and UR-FUNNY. The probability of masking and AP ratio were 15% and 50%, respectively.

5. Results and analysis

5.1. Comparative results for MSA

5.1.1. Results for CMU-MOSEI

The results of CMU-MOSEI sentiment analysis are presented in Table 2. AOBERT achieved the best performance and outperformed the baselines across all the metrics. It exhibited improvements of 2.3 and 0.04 points in A7 and the MAE, respectively. These results indicated that the proposed model can provide more accurate predictions than the baseline models.

The results of the proposed model were better than those of models with intricate fusion mechanisms, such as TFN [2] and LMF [32]. In addition, the single-stream transformer outperformed RAVEN [8] and MulT [7], which utilize vanilla attention and cross-transformer. Thus, we demonstrated that a single-stream transformer is more effective. MMIM [33] and HyCon [34] use a contrastive learning method for multimodal areas. The contrastive learning method involves self-supervised representation learning, which has led to major advances in vision research area. Using contrastive learning, the model can achieve the same performance that is achieved via data augmentation. However, AOBERT achieved a better score without using contrastive learning.

5.1.2. Results for CMU-MOSI'

The sentiment analysis results for CMU-MOSI are presented in Table 3. AOBERT had the best score for all the metrics except MAE and Corr. It outperformed joint representation models such as MFN [35], MCTN [5], BC-LSTM [36], and MARN [27] based on the RNN. This indicated that AOBERT can address the long-term dependency problem using attention mechanism. ICCN [37] has a similar method to the proposed model. It first extracts features from the Speech and Vision modalities and then fuses each with Text embeddings to obtain two outer products. The outer results are fed to a Canonical Correlation Analysis and the corresponding output is used for prediction. However, AOBERT achieved a higher score than ICCN, indicating that it has a more effective fusion method.

However, the proposed model did not outperform the state-of-the-art model, MISA, with regard to the MAE. This is because the CMU-MOSI dataset was not large enough for training the proposed large model. In general, data augmentation is used to increase the size of a dataset for solving the problem of insufficient data size. Although there are data augmentation methods for unimodal data such as Text or Vision, to the best of our knowledge, there are no approaches for multimodal data. Contrastive learning, which has the same effect as data augmentation, can solve the aforementioned problem will be investigated in a future study.

5.2. Comparative results for multimodal emotion detection

5.2.1. Results for UR-FUNNY

Similar trends were observed for MHD (Table 4). Humor detection is highly sensitive to the idiosyncratic characteristics of different modalities. AOBERT outperformed existing models, such as MISA, state-of-the-art model that considers both inter-modality and intra-modality. Additionally, it was verified that AOBERT covers both inter-modality and intra-modality.

5.2.2. Results of CMU-MOSEI

Table 5 presents the results of CMU-MOSEI emotion detection, where A2 and F1 were used as the metrics. Emotion detection was conducted in the form of an independent model for each emotion. In Graph-MFN [20] and MTMM-ES [21], A2 was used with the weighted method. However, TBJE [38] and AOBERT used standard methods. For a fair comparison, we excluded the A2 accuracy of Graph-MFN and MTMM-ES. TBJE-2 and TBJE-3 were used for TBJE, where the number refers to the number of modalities used (the TBJE-2 model uses Text and Speech modalities). Because TBJE-2 achieved a better score than TBJE-3, we present TBJE-2 and TBJE-3 together for a fair comparison.

AOBERT achieved the highest performance for all emotions except 'Fear' and 'Surprise' and obtains a good A2 score for 'Fear'. TBJE utilizes a transformer architecture and relies on a glimpse layer to encode one or more modalities jointly. AOBERT outperformed TBJE, indicating that the proposed model can effectively encode modalities and generate a joint representation. The distribution of emotion labels in CMU-MOSEI was unbalanced. In particular, only approximately 2000 data corresponded to 'Fear' and 'Surprise' out of a total of 23453 data. Thus, the results indicated the effectiveness of AOBERT for balanced data.

5.3. Ablation study

We performed ablation studies to investigate the characteristics of each modality and the influence of each component, such as MMLM and self-attention. (-) denotes that the experiment was conducted except for a pair of modalities or an element in AOBERT as A2 and, MAE metrics.

We removed the pairs T, V', and S' individually to determine the influence of the modality. In addition, we deleted components, such as the MMLM, and self-attention layer to examine the effects. The results of the ablation study are presented in Table 6.

5.3.1. Combinations of modalities

An investigation was also conducted to determine the effects of removing each pair of modalities on the performance, for CMU-MOSI, CMU-MOSEI, and UR-FUNNY. The results are presented in the middle part of Table 6.

Among all the cases, the performance was the best when all modalities were used. For the MAE, the most significant difference occurred when one modality was removed. The MAE results indicated a significant reduction in the prediction accuracy when one modality was removed. When the V' or S' pair was removed, the performance decline was more significant than that when the T was removed. Because V' and S' also had text utterances, the performance hardly declined when T was removed. The lowest scores were obtained when V' was ejected, indicating that Vision contributes significantly to the prediction.

5.3.2. Self-attention in classifier

Using the self-attention layer in the classifier, we obtained a meaningful joint representation. When the self-attention layer is removed, AOBERT predicts sentiment or humor using only one fully connected layer.

As shown in Table 6, the overall performance declined when the self-attention layer is removed from the model. The A2 score did not decrease significantly for CMU-MOSEI, and the MAE increased. The

model only makes rough predictions and cannot make elaborate predictions with only fully connected layer. The results indicated that the self-attention layer guarantees a more precise joint representation for predicting actual labels.

5.3.3. Effect of MMLM

We can learn joint representations using MMLM to predict the masked tokens. To verify the effect of the MMLM, we performed only the AP task in the experiments.

As shown in Table 6, the most significant performance degradation was observed for CMU-MOSI and UR-FUNNY when MMLM was removed. For CMU-MOSEI, an appropriate joint representation was learned to predict the sentiment or emotion with only AP owing to the large size of dataset. However, the results for CMU-MOSI and UR-FUNNY were worse because of the relatively small sizes of datasets. Therefore, MMLM is a critical task in AOBERT training.

5.3.4. Effect of AP

AOBERT can learn the relationship between modalities by performing AP. A comparison experiment was conducted on the AP ratio to determine the effect of AP on the CMU-MOSEI dataset. When the AP ratio was 1.0, there were no "UnPair" samples, indicating that the AP was not performed.

Table 7 presents the results of a comparison experiment to find the optimum AP ratio (0.3-1.0, steps of 0.1), which was determined to be 0.5. The differences in the results between the cases where the AP ratio was optimal and the other cases, (except for 1.0), were small. When the AP ratio was 1.0, the MAE exceeded 0.6. This indicated that the model cannot capture the relationship between modalities without AP, which suggests that AOBERT can capture inter-modality using AP.

5.3.5. Fusion method

Multimodal research has been performed by fusing the data at the input or prediction stage, e.g., Early-Fusion and Late-Fusion. Similarly, AOBERT uses the fusion method at Fusion Gate and Prediction Stage. However, in contrast to other multimodal research, AOBERT can interpret intact information of modalities for learning representations and final prediction using these fusion methods. Early-Fusion integrates the information of different modalities in the input stage, so that the Fusion Gate can be replaced. By contrast, Late-Fusion represents each modality as an independent model and applies a majority-voting or weighted average approach to the model's results. Therefore, Late-Fusion can substitute Prediction Stage.

To validate our fusion method, we adopted Early-Fusion and Late-Fusion instead of Fusion Gate and Prediction Stage. In Table 8, FG and PS denote Fusion Gate and Prediction Stage, respectively. We experimented "Early-Fusion + PS" and "FG + Late-Fusion" on the CMU-MOSEI dataset. Neither fusion method differed significantly from the A2 method. However, there were many declines of approximately six points in A7, which is a more sophisticated metric. The results indicated that AOBERT fusion method helps to make more accurate predictions owing to the intact unique features of the modality.

5.4. Qualitative analysis

We examined the advantages of MSA by sampling the results of the trained model for the CMU-MOSI dataset. We evaluated AOBERT with regard to the A2 metric using three modalities and only Text.

Table 9 presents sentences and prediction results of the models, and Fig. 4 shows screenshots of the speakers for each sentence. Speech is not included due to the limitation of the paper. (a), (b) are cases where both models predicted sentiments incorrectly. (c), (d) are dominant examples in which the multi-modal model was used. (e), (f) are cases where better performance was achieved using only Text.

Sentences (a) and (b) appear to be positive and negative, respectively. In addition, the individuals spoke within a solid tone and gestures. However, in reality, the label has a value of zero or a low value. Thus, the A2 metric excludes zero labels because the example indicated

Fig. 4. Examples from the qualitative analysis.

Table 1
Datasets used.

Datasets	CMU-MOSI	CMU-MOSEI	UR-FUNNY
outputs	Sentiment	Sentiment, Emotion	Humor
training	1,284	16,216	7,614
validation	229	1,871	980
test	686	4,654	994

Table 2
Performance comparison with previous models for CMU-MOSEI

Model	A2(↑)	A7(↑)	MAE(↓)	Corr(↑)	F1(↑)
Graph-MFN [20]	76.9 / -	45.0	0.710	0.540	77.0 / -
RAVEN [8]	79.1 / -	50.0	0.614	0.662	79.5 / -
MCTN [5]	79.8 / -	49.6	0.609	0.670	80.6 / -
CIA [42]	80.4 / -	50.1	0.680	0.590	78.2 / -
MulT [7]	- / 82.5	51.8	0.580	0.703	- / 82.3
TFN (B) [2]	- / 82.5	50.2	0.593	0.700	- / 82.1
LMF (B) [32]	- / 82.0	48.0	0.623	0.677	- / 82.1
MFM (B) [35]	- / 84.4	51.3	0.568	0.717	- / 84.3
ICCN (B) [37]	- / 84.2	51.6	0.565	0.713	- / 84.2
MISA (B) [26]	83.6 / 85.5	52.2	0.555	0.756	83.8 / 85.3
MMIM (B) [33]	82.2 / 86.0	54.2	0.526	0.772	82.7 / **85.9**
HyCon (B) [34]	- / 85.4	52.8	0.601	**0.778**	- / 85.6
AOBERT (B)	**84.9 / 86.2**	**54.5**	**0.515**	0.763	**85.0 / 85.9**

Table 3
Performance comparison with previous models for CMU-MOSI

Model	A2(↑)	A7(↑)	MAE(↓)	Corr(↑)	F1(↑)
BC-LSTM [36]	73.9 / -	28.7	1.079	0.581	73.9 / -
MV-LSTM [43]	73.9 / -	33.2	1.019	0.601	74.0 / -
TFN [2]	73.9 / -	32.1	0.970	0.633	73.4 / -
MARN [27]	77.1 / -	34.7	0.968	0.625	77.0 / -
MFN [35]	77.4 / -	34.1	0.965	0.632	77.3 / -
LMF [32]	76.4 / -	32.8	0.912	0.668	75.7 / -
MFM [44]	78.1 / -	36.2	0.951	0.662	78.1 / -
RAVEN [8]	76.6 / -	33.2	0.915	0.691	76.6 / -
RMFN [45]	78.0 / -	38.3	0.922	0.681	78.0 / -
MCTN [5]	79.1 / -	35.6	0.909	0.676	79.1 / -
CIA [42]	79.8 / -	38.9	0.914	0.689	- / 79.5
MulT [7]	- / 83.0	40.0	0.871	0.698	- / 82.8
TFN (B) [2]	- / 80.8	34.9	0.901	0.698	- / 80.7
LMF (B) [32]	- / 82.5	33.2	0.917	0.695	- / 82.4
MFM (B) [42]	- / 81.7	35.4	0.877	0.706	- / 81.6
ICCN (B) [37]	- / 83.0	39.0	0.860	0.710	- / 83.0
MISA (B) [26]	81.8 / 83.4	42.3	0.783	0.761	81.7 / 83.6
MMIM (B)[33]	84.1 / **86.1**	**46.7**	**0.700**	**0.800**	84.0 / 86.0
HyCon (B) [34]	- / 85.2	46.6	0.713	0.790	- / 85.1
AOBERT (B)	**85.2** / 85.6	40.2	0.856	0.700	**85.4 / 86.4**

that zero labels negatively affect the results.

(c) and (d) indicate that facial expressions and voice influence sentiment prediction. There are no particular points for predicting sentiment when focusing only on sentences. However, the model confidently predicted a negative sentiment because the voice had an angry tone for (c), whereas for (d), the model predicted a positive sentiment using facial expressions. The results indicated predictions with Vision and Speech are more accurate than those using only Text.

The last two examples indicate that the multimodal approach is sometimes unhelpful for accurate predictions. Although both facial expressions and voice tones included in videos, clues concerning sentiments appear in the Text. Therefore, the results change when all modalities are used together. Because the changes in the prediction values are insignificant, a more specific joint representation is a potential method for solving this problem.

Table 4
Performance comparison with previous models for UR-FUNNY.

Models	A2(↑)
C-MFN [23]	64.47
TFN [2]	64.71
LMF [32]	65.16
LMF (B) [32]	67.53
TFN (B) [2]	68.57
MISA (G) [26]	68.60
MISA (B) [26]	70.61
AOBERT (B)	**70.82**

Table 5
Performance comparison with previous models for CMU-MOSEI emotion detection

Emotion	Happy		Sad		Anger	
Metric	**A2(↑)**	**F1(↑)**	**A2(↑)**	**F1(↑)**	**A2(↑)**	**F1(↑)**
Graph-MFN [20]	-	66.3	-	66.9	-	72.8
MTMM-ES [21]	-	67.0	-	**72.4**		75.9
TBJE-2 [38]	66.0	65.5	73.9	67.9	81.9	76.0
TBJE-3	65.0	64.0	72.0	67.9	81.6	74.7
AOBERT	**68.4**	**68.2**	**74.0**	71.4	**82.5**	**77.4**
Emotion	**Fear**		**Disgust**		**Surprise**	
Metric	**A2(↑)**	**F1(↑)**	**A2(↑)**	**F1(↑)**	**A2(↑)**	**F1(↑)**
Graph-MFN	-	**89.9**	-	76.6	-	85.5
MTMM-ES		87.9		81.9		86.0
TBJE-2	89.2	87.2	86.5	84.5	**90.6**	**86.1**
TBJE-3	89.1	84.0	85.9	83.6	90.5	**86.1**
AOBERT	**90.6**	86.1	**87.2**	**85.3**	89.7	**86.1**

Table 6
Results of the ablation study

Model	CMU-MOSI		CMU-MOSEI		UR-FUNNY
	A2	MAE	A2	MAE	
AOBERT	**83.0/ 84.3**	**0.825**	**84.9/ 86.2**	**0.515**	**70.81**
(-) Text T	79.9/ 81.9	0.901	83.0/ 83.0	0.593	69.28
(-) Vision V'	77.7/ 77.7	1.201	82.5/ 85.2	0.619	68.47
(-) Speech S'	78.6/ 79.1	0.996	82.6/ 84.0	0.588	69.30
(-) MMLM	76.9/ 78.2	1.022	83.2/ 84.0	0.554	67.85
(-) self-attention	78.2/ 79.2	0.972	83.1/ 83.2	0.625	68.78

Table 7
Effect of the AP ratio

AP ratio	A2(↑)	MAE(↓)	F1(↑)
0.3	84.1/85.9	0.559	84.3/85.6
0.4	84.1/85.5	0.532	84.3/85.3
0.5	**84.9/86.2**	**0.515**	**85.0/85.9**
0.6	83.9/85.0	0.567	84.0/84.7
0.7	84.1/85.9	0.552	84.4/85.8
0.8	84.2/85.5	0.537	84.3/85.3
0.9	84.7/85.6	0.539	84.8/85.4
1.0	83.4/84.6	0.613	83.4/84.3

Table 8
Fusion method

Method	CMU-MOSEI A2	A7	MAE	Corr	F1
FG+PS	**84.9 / 86.2**	**54.5**	**0.515**	**0.763**	**85.0 / 85.9**
Early-Fusion + PS	83.6/85.3	48.2	0.586	0.754	83.6/85.5
FG + Late-Fusion	83.9/85.6	48.9	0.577	0.753	83.9/85.8

Table 9
Sentences and predictions for Fig. 4

Index	Sentence	True	T	T, V', S'
(a)	I thought this movie was gonna be really good because matt damon was	– 0.75	1.9	2.4
(b)	bad i don't think it's like the worst movie I have even seen just	0	– 1.3	– 0.4
(c)	on oh god why did i watch that	– **2.8**	0.3	– **0.8**
(d)	if you don't have kids you may wanna consider seeing it	**1.8**	– 0.1	**2.0**
(e)	I'm not really in a big hurry at my age to go see animated	– **0.2**	– **0.1**	0.2
(f)	But it's not always quite explained you know like everything	– **0.6**	– **0.5**	0.4

6. Conclusion

MSA utilizes Text and various modalities, such as Vision or Speech to predict sentiment. The modality fusion mechanism is essential since each modality has different characteristics. Thus, we propose the AOBERT method, which uses a single-stream transformer. Text, Vision, and Speech pass through one network and are processed by two pre-training tasks: MMLM and AP. In addition, we propose sophisticated prediction using self-attention in the classification stage.

To verify the performance of the proposed model, we conducted experiments on three datasets: CMU-MOSI, MOSEI, and UR-FUNNY for MSA, Emotion Detection, and MHD. AOBERT outperformed previously proposed models, including the state-of-the-art model MISA. In addition, we performed ablation studies on the combination of modalities, the influence of MMLM, AP, and self-attention to prove the necessity of each component.

Sentiment analysis has diverse polarities, such as ambivalence [39]. Detecting ambivalence is an important task, but there is no multimodal dataset for ambivalence polarity. To detect sentiments in real-world conversations, the MELD dataset [40] can be used. It contains more than 1400 dialogues and 13000 utterances from the Friends TV series for emotion recognition. Using reinforcement learning and domain knowledge, real-time video emotion recognition was performed for the MELD dataset [41]. Therefore, we consider that AOBERT can be improved to adapt to diverse polarities and recent efforts. We can also fine-tune AOBERT for other multimodal tasks, such as video questioning and answering, using joint representations derived from the model.

CRediT authorship contribution statement

Kyeonghun Kim: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Sanghyun Park:** Validation, Resources, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (IITP-2017-0-00477, (SW starlab) Research and development of the high performance in-memory distributed DBMS based on flash memory storage in an IoT environment).

References

[1] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective computing and sentiment analysis, in: A practical guide to sentiment analysis, Springer, Cham., 2017, pp. 1–10, https://doi.org/10.1007/978-3-319-55394-8_1.
[2] M.Chen Zadeh, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1103–1114, https://doi.org/10.18653/v1/D17-1115.
[3] J.Vepa Kumar, Gated mechanism for attention based multimodal sentiment analysis, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 4477–4481, https://doi.org/10.1109/ICASSP40776.2020.9053012.
[4] A.Sardana Shenoy, N. Graphics, Multilogue-net: a context aware rnn for multi-modal emotion detection and sentiment analysis in conversation, ACL (2020) 19–28, https://doi.org/10.18653/v1/2020.challengehml-1.3.
[5] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, B. Poczos, Found in translation: learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence 33, 2019, pp. 6892–6899, https://doi.org/10.1609/aaai.v33i01.33016892.
[6] N.Shazeer Vaswani, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, In Adv. Neural Inf. Process. Syst. (2017) 5998–6008, https://doi.org/10.5555/3295222.3295349.
[7] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, pp. 6558–6569, https://doi.org/10.18653/v1/P19-1656.
[8] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.-P. Morency, Words can shift: dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence 33, 2019, pp. 7216–7223, https://doi.org/10.1609/aaai.v33i01.33017216.
[9] A. Baevski, W.N. Hsu, Q. Xu, A. Babu, J. Gu, M. Auli, Data2vec: a general framework for self-supervised learning in speech, vision and language, arXiv preprint (2022), https://doi.org/10.48550/arXiv.2202.03555 arXiv:2202.03555.
[10] J. Ao, R. Wang, L. Zhou, S. Liu, S. Ren, Y. Wu, F. Wei, Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing, arXiv preprint (2021), https://doi.org/10.48550/arXiv:2110.07205 arXiv:2110.07205.
[11] X. Chen, X. Song, L. Jing, S. Li, L. Hu, L Nie, Multimodal dialog systems with dual knowledge-enhanced generative pretrained language model, arXiv preprint (2022), https://doi.org/10.48550/arXiv:2207.07934 arXiv:2207.07934.
[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota,, 2019, pp. 4171–4186, https://doi.org/10.18653/v1/N19-1423.
[13] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, D. Amodei, Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33 (2020) 1877–1901, https://doi.org/10.48550/arXiv.2005.14165.
[14] W. Zhang, X. Li, Y. Deng, L Bing, W. Lam, A survey on aspect-based sentiment analysis: tasks, methods, and challenges, arXiv preprint (2022), https://doi.org/10.48550/arXiv:2203.01054 arXiv:2203.01054.
[15] D. Araci, Finbert: Financial sentiment analysis with pre-trained language models, arXiv preprint (2019), https://doi.org/10.48550/arXiv:1908.10063 arXiv:1908.10063.
[16] H. Grissette, E.H Nfaoui, Deep associative learning approach for bio-medical sentiment analysis utilizing unsupervised representation from large-scale patients' narratives, Person. Ubiquit. Comput. (2021) 1–15, https://doi.org/10.1007/s00779-021-01595-4.
[17] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: scalable multimodal fusion for the continuous interpretation of semantics and sentics, in: 2013 IEEE symposium on computational intelligence for human-like intelligence (CIHLI), 2013, pp. 108–117, https://doi.org/10.1109/CIHLI.2013.6613272. April.

[18] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th international conference on data mining (ICDM), IEEE, 2016, pp. 439–448, https://doi.org/10.1109/ICDM.2016.0055.

[19] D.Gopinath Nojavanasghari, J. Koushik, T. Baltrušaitis, L.-P. Morency, Deep multimodal fusion for persuasiveness prediction, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 284–288, https://doi.org/10.1145/2993148.2993176.

[20] B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, 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), 2018, pp. 2236–2246, https://doi.org/10.18653/v1/P18-1208.

[21] M.S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi-task learning for multi-modal emotion recognition and sentiment analysis, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1, 2019, pp. 370–379, https://doi.org/10.18653/v1/N19-1034.

[22] Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, arXiv preprint arXiv:1606.06259. https://doi.org/10.48550/arXiv.1606.06259.

[23] M.K. Hasan, W. Rahman, A. Bagher Zadeh, J. Zhong, M.I. Tanveer, L.-P. Morency, M.E. Hoque, UR-FUNNY: a multimodal language dataset for understanding humor, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2046–2056, https://doi.org/10.18653/v1/D19-1211.

[24] A.Perelygin Socher, J. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.

[25] P. Ekman, W.V. Freisen, S. Ancoli, Facial signs of emotional experience, J. Personal. Soc. Psychol. 39 (6) (1980) 1125–1134, https://doi.org/10.1037/h0077722.

[26] R.Zimmermann Hazarika, S. Poria, Misa: Modality-invariant and specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131, https://doi.org/10.1145/3394171.3413678.

[27] P. Zadeh, P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 5642–5649, https://doi.org/10.48550/arXiv.1802.00923.

[28] J. Pennington, R. Socher, C.D. Manning, Glove: global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543, https://doi.org/10.3115/v1/D14-1162.

[29] B. Moeslund, A. Hilton, V. Krüger, L. Sigal, Visual Analysis of Humans, Springer, 2011, https://doi.org/10.1007/978-0-85729-997-0.

[30] A.Zadeh Baltrusaitis, Y.C. Lim, L.-P. Morency, Openface 2.0: facial behavior analysis toolkit, in: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 59–66, https://doi.org/10.1109/FG.2018.00019.

[31] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep—a collaborative voice analysis repository for speech technologies, in: 2014 IEEE international conference on acoustics, speech and signal processing (icassp), IEEE, 2014, pp. 960–964, https://doi.org/10.1109/ICASSP.2014.6853739.

[32] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A.B. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2247–2256, https://doi.org/10.18653/v1/P18-1209.

[33] W. Han, H. Chen, S Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, arXiv preprint (2021), https://doi.org/10.48550/arXiv.2109.00412 arXiv:2109.00412.

[34] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022) 1, https://doi.org/10.1109/TAFFC.2022.3172360. -1.

[35] P.P.Liang Zadeh, N. Mazumder, S. Poria, E. Cambria, L.P. Morency, Memory fusion network for multi-view sequential learning, Proc. AAAI Conf. Artif. Intell. 32 (2018) 5634–5641, https://doi.org/10.48550/arXiv.1802.00927.

[36] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, 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), 2017, pp. 873–883, https://doi.org/10.18653/v1/P17-1081.

[37] P.Sarma Sun, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence 34, 2020, pp. 8992–8999, https://doi.org/10.1609/aaai.v34i05.6431.

[38] J.-B. Delbrouck, N. Tits, M. Brousmiche, S. Dupont, A transformer-based joint-encoding for emotion recognition and sentiment analysis, ACL 2020 (2020) 1–7, https://doi.org/10.18653/v1/2020.challengehml-1.1.

[39] Z. Wang, S.B. Ho, E. Cambria, Multi-level fine-scaled sentiment sensing with ambivalence handling, Int. J. Uncertain. Fuzzin. Knowl.-Based Syst. 28 No (.04) (2020) 683–697, https://doi.org/10.1142/S0218488520500294.

[40] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R.Meld Mihalcea, A multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint (2018), https://doi.org/10.48550/arXiv.1810.02508 arXiv:1810.02508.

[41] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-Time Video Emotion Recognition based on Reinforcement Learning and Domain Knowledge, 32, IEEE Transactions on Circuits and Systems for Video Technology, 2021, pp. 1034–1047, https://doi.org/10.1109/TCSVT.2021.3072412.

[42] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, Pushpak Bhattacharyya, Context-aware interactive attention for multi-modal sentiment and emotion analysis, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5647–5657, https://doi.org/10.18653/v1/D19-1566.

[43] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, Roland Goecke, Extending long short-term memory for multi-view structured learning, in: Computer Vision - ECCV 2016 - 14th European Conference, Proceedings 9911, Springer, Amsterdam, The Netherlands, 2016, pp. 338–353, https://doi.org/10.1007/978-3-319-46478- 7_21. Part VII (Lecture Notes in Computer Science.

[44] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, Ruslan Salakhutdinov, Learning factorized multimodal representations, in: 7th International Conference on Learning Representations, ICLR, OpenReview.net, New Orleans, LA, USA, 2019. https://openreview.net/forum?id=rygqqsA9KX.

[45] Paul Pu Liang, Ziyin Liu, Amir Zadeh, Louis-Philippe Morency, Multimodal language analysis with recurrent multistage fusion, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 150–161, https://doi.org/10.18653/v1/d18-1014.