Shared-Private Memory Networks For Multimodal Sentiment Analysis

Xianbing Zhao , Yinxin Chen, Sicen Liu , and Buzhou Tang, *Member, IEEE*

Abstract—Text, visual, and acoustic are usually complementary in the Multimodal Sentiment Analysis (MSA) task. However, current methods primarily concern shared representations while neglecting the critical private aspects of data within individual modalities. In this work, we propose shared-private memory networks based on the recent advances in the attention mechanism, called SPMN, to decouple multimodal representation from shared and private perspectives. It contains three components: a) a shared memory to learn the shared representations of multimodal data; b) three private memories to learn the private representations of individual modalities, respectively; c) and adaptive fusion gates to fuse multimodal private and shared representations. To evaluate the effectiveness of SPMN, we integrate it into different pre-trained language representation models, such as BERT and XLNET, and conduct experiments on two public datasets, CMU-MOSI and CMU-MOSEI. Experimental results indicate that the performances of pre-trained language representation models are significantly improved because of SPMN and demonstrate the superiority of our model compared to the state-of-the-art methods. SPMN's source code is publicly available at: https://github.com/xiaobaicaihhh/SPMN.

Index Terms—Multimodal sentiment analysis, shared-private memory netwoks, adaptive fusion gate, BERT

◆

1 INTRODUCTION

MULTIMODAL Sentiment Analysis (MSA) has attracted more and more attention [1], [2], [3], [4], [5], [6] in recent years. Multimodal representation and fusion are two challenges of MSA. Lots of studies have been proposed to tackle them. For multimodal representation, researchers [7], [8], [9], [10] adopt different neural networks to learn specific representations of multimodal data and gain considerable performance improvements. In the perspective of multimodal fusion [11], [12], [13], [14], crossmodal attention [1], additive attention [4] and gating mechanism [15] are usually used to fuse multimodal information. Most of the existing MSA methods only focus on shared multimodal representations

- *Xianbing Zhao, Yinxin Chen, and Sicen Liu are with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong Province 518055, China. E-mail: zhaoxianbing_hitsz@163.com, cyxhelloo@gmail.com, liusicen_cs@outlook.com.*
- *Buzhou Tang is with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong Province 518055, China, and also with Pengcheng Laboratory, Guangdong Province 518066, China. E-mail: tangbuzhou@hit.edu.cn.*

and fusion, and then conduct them from a shared perspective. However, sentiment attitude often involves shared and private perspectives in multimodal sentiment analysis tasks.

As shown in Fig. 1, in the CMU-MOSEI and CMU-MOSI datasets, the author recommends a movie with facial and verbal expressions. The disappointed facial expressions of visual modality and the "gosh" of text modality are shared factors, while the "squinted" of visual modality and the "bad movie" of text modality are private factors, respectively. Both shared and private factors have an important influence on multimodal sentiment analysis. Therefore, the multimodal representation's accurate modeling must consider both the shared and private aspects.

Recently, pre-trained language representation models [16], [17] based on Transformer [18] have been extended for multimodal fusion and achieved state-of-the-art performance. The representative method is MAG-BERT [15], which introduced a multimodal adaptive gate to fuse the visual and acoustic representations at the specific layer of BERT. It is a typical fusion method without considering multimodal representations from shared and private perspectives.

Motivated by recent advances in domain adaptation [9], [15], we propose shared-private memory networks that can learn both the shared and the private representations for each modality, called SPMN. SPMN extends the well-known family of BERT [16] by introducing separate shared and private memory networks and adaptive fusion gates. In detail, we design several memory matrices to store the modal shared and private information and employ it to perform memory query and memory response for each modality. For memory query, we generate response vectors by weighting the queried memory items. As a result, SPMN learns decoupled representations (i.e., shared and private representations) for each modality by utilizing different memory networks.

With respect to learn shared representations, each modality queries from the same memory items and obtains different

(a) Visual and text modality
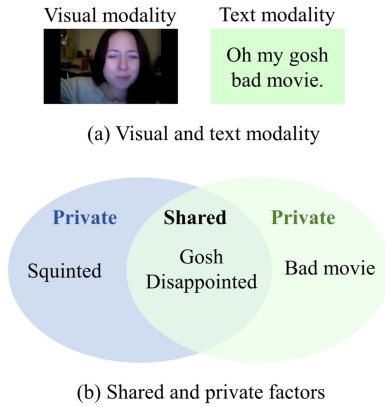

(b) Shared and private factors

Fig. 1. (a) An example of bimodal data. Modality V is a facial expression of a person, and modality T represents a text, where the person describe a movie. (b) Only some of the factors are shared by both modalities in shared $V \cap T$. Other factors are private to individual modalities, grouped in separate Private spaces. By definition, the three spaces are disentangled from each other.

response vectors. Shared memory networks reduce the modality gaps when the model performs multimodal interactions. Though the different modalities reside in separate representation spaces, they share the same sentiment attitude. The shared memory help captures cross-modal underlying commonalities and generates shared representation across modality. Regarding modeling private representations, each modality queries from private memory items and obtains individual response vectors, which are closely related to the modality. Each modality holds distinctive features. Such idiosyncratic features are lowly correlated with other modalities for cross-modal interaction and are often categorized as noise. Nevertheless, they may be helpful for sentiment analysis. As shown in Fig. 1, such as a speaker uses ridicule to express negative sentiment, while the text modality appears as neutral sentiment. Therefore, the model considers the private aspects of individual modalities and what those modalities share. Shared and private representations are complementary. Finally, SPMN adopts adaptive fusion gates to fuse shared and private representations and integrates the final representation into BERT's backbone. The novel contributions of our work can be summarized as follows:

- We design shared-private memory networks to learn the shared and private representations of complementary multimodal data, respectively.
- We design shared and private adaptive fusion gates to extract text-related information from shared and private representations for pre-trained language representation model fine-tuning.
- Extensive experiments on two benchmark datasets CMU-MOSEI and CMU-MOSI show that our method outperforms state-of-the-art methods.

2 RELATED WORK

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis mainly involves three modalities, text, visual, and acoustic, and employs these modalities to comprehend varied human sentiment. According to learning the multimodal unified representations, we roughly divide existing methods into two groups: fusing

methods through task loss back-propagation or geometric manipulation in the embedding space.

The former accomplishes back-propagated gradients via tuning the parameters from the specific task loss. Specifically, Zadeh et al. [11] proposed a Tensor Fusion Network (TFN) to capture unimodal, bimodal, and trimodal interactive information. Following the TFN, Liu et al. [12] designed a Low-rank Multimodal Fusion (LMF) to improve the efficiency of multiple tensor fusion. Based on the time-dependent multimodal data, Zadeh et al. [19] and Wang et al. [20] proposed Memory Fusion Networks (MFN) and Recurrent Attended Variation Embedding Network (RAVEN) from different perspectives. MFN employed multi-view sequential learning via integrating both view-specific and cross-view information. RAVEN designed a fine-grained architecture to shift verbal token representations based on nonverbal cues dynamically. With the development of Transformer [18], multimodal sentiment analysis achieved outstanding performance via exploiting the attention mechanism. Tsai et al. [1] proposed a Multimodal Transformer (MuLT), which employs cross-modal attention to capture intra-modal information. Lv et al. [4] designed a Progressive Modality Reinforcement (PMR), which contains a message hub to reinforce each modality via cross-modal interaction. Rahman et al. [15] proposed an adaptive gate to integrate multimodal information into the BERT backbone, which achieves state-of-the-art performance due to the efficiency of the pre-trained language representation model.

The latter branch aims to rectify unimodal or multimodal representations via exploiting fine-grained geometric operations in embedding space. To be specific, Tsai et al. [7] proposed a Multimodal Factorization Model (MFM) to model multimodal representation via exploiting multimodal discriminative and modality-specific generative factors. Hazarika et al. [9] proposed Modality-Invariant and -Specific Representations (MISA) to model modality-invariant and modality-specific representation by utilizing multiple different losses. After that, plentiful modality-specific representations were explored and applied to the multimodal sentiment analysis task. More specifically, Yu et al. [10] proposed a self-supervised learning method to generate unimodal supervision. Han et al. [21] proposed MultiModal InfoMax (MMIM) to maximize the mutual information between bimodality. Han et al. [22] also proposed a Bi-Bimodal Fusion Network (BBFN) to perform fusion and separation on pairwise modality representations.

2.2 Pre-Trained Language Representation Model

Transformer is a sequence-to-sequence modal that translates one language to another in NMT (Neural Machine Translation) [18] task. Recently, pre-trained language representation models based on Transformer [18] have achieved superior performance on several NLP (Natural Language Processing) tasks. BERT (Bidirectional Encoder Representations Transformers) [16] is a typical pre-trained language representation model and achieved remarkable performance in multiple downstream tasks. In addition, different pre-trained language representation models give multiple new contextual representations through building an autoregressive model, such as XLNET [23], RoBerta [24], DeBerta [25], and Electra [26]. Currently, there are two
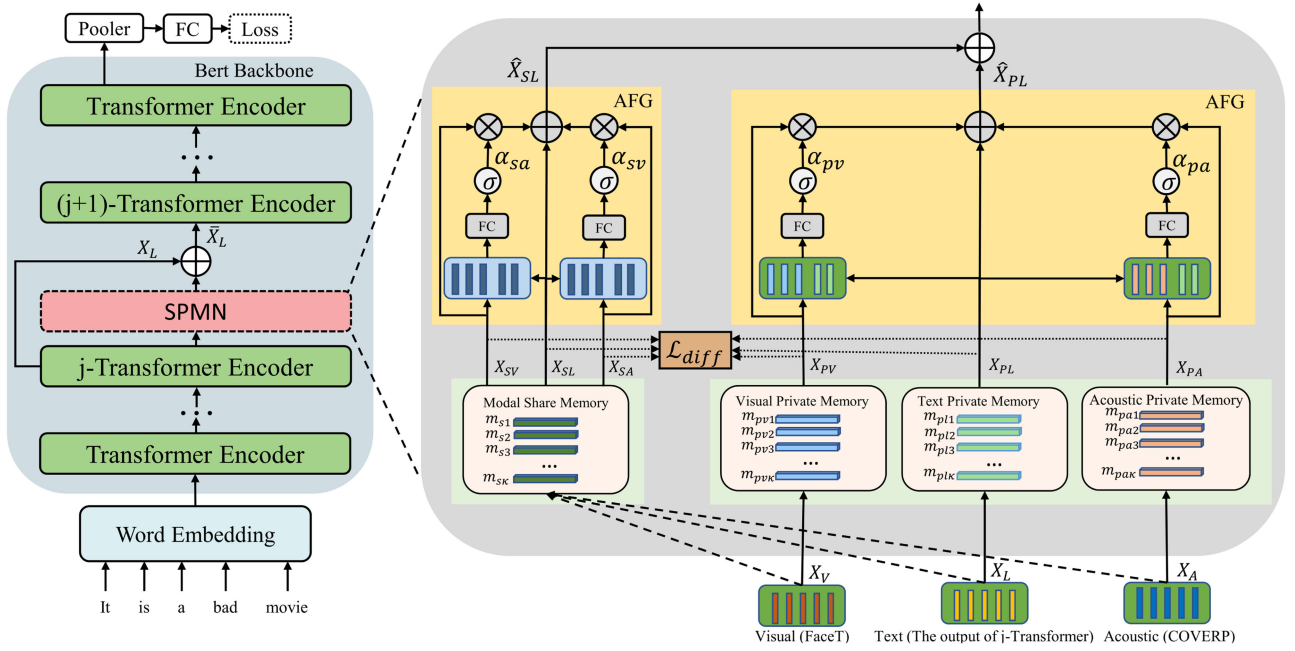
Fig. 2. *Illustration of SPMN.* It consists of a shared memory, three private memories, two adaptive fusion gates, and a BERT's backbone. We adopt a difference loss to reinforce the shared and private representation. Shared and private representations are fused by Adaptive Fusion Gate (AFG) respectively, and the result is integrated into BERT's backbone.

different ways to employ these pre-trained language representation models. The first way is to utilize the pre-trained language representation model as a feature extraction module [9], [10]. The second way is to integrate multimodal information into specific layers of the BERT backbone [15]. In this work, we adopt the second way and fine-tuning the pre-trained language representation model for multimodal sentiment analysis tasks.

2.3 Memory Networks

Memory networks store information and read the relevant contents from the external memory, which consists of multiple learnable vectors. Memory networks have recently been proposed to solve various problems in Nature Language Processing [27], [28], [29], [30] and Computer Vision [31], [32], [33], [34] tasks. The Key-Value Memory Networks [35] exploits a key-value structured memory to enhance the encoder-decoder framework for radiology report generation.

Our work is inspired by the disentangled multimodal representation [9], [36] strategy. MISA [9] learns modality-invariant and modality-specific representations for each modality by multiple combination of losses and encoder networks. DMVAE [36] introduces a disentangled multimodal variational auto-encoder that utilizes disentangled VAE (Variational Autoencoder) strategy to separate the private and shared latent spaces of multiple modalities. Recent work has introduced memory networks on many fields [31], [33], [34], [35], and obtained enhancements. Our approach is the first to introduce the idea of memory networks to the multimodal sentiment analysis task with three modalities text, visual and acoutsic. Different from MISA [9] and DMAVE [36], we employ memory networks to construct decoupled multimodal representations. To utilize the external memory networks for our purposes, we introduce a novel shared-private memory networks to learn two distinct

representations for each modality. We exploits backward gradients to update the learnable shared and private memory items. Then the model reads shared and private representation from the learnable shared and private memory items, respectively. Then we design adaptive fusion gate to fuse multiple decoupled modality representations. Finally, we also explore the impact of the representation of the pre-trained language representation model on SPMN.

3 PROPOSED APPROACH

In this section, we elaborate the overall architecture of SPMN and its main components, as illustrated in Fig. 2. Concretely, we first introduce the pre-trained language representation model (i.e., BERT) in Section 3.2 and the memory networks construction in Section 3.3. Afterwards, we connect varied memory networks to construct the shared-private memory networks in Section 3.4 and present two adaptive fusion gates to integrate multimodal representations in Section 3.5.

3.1 Model Overview

The inputs to the model are the representation of three modalities $X_L \in R^{T \times d_l}$, $X_V \in R^{T \times d_v}$, and $X_A \in R^{T \times d_a}$. Following [15], we use a fully connected layer to process the input sequences $X_L \in R^{T \times d_l}$, $X_V \in R^{T \times d_v}$, and $X_A \in R^{T \times d_a}$ to obtain representations of unified feature dimensions (i.e., $X_L \in R^{T \times d}$, $X_V \in R^{T \times d}$, and $X_A \in R^{T \times d}$), where d represents feature dimension (i.e., BERT feature dimension 768). The inputs pass through the fully connected layer to keep the feature dimension consistent. Note that the input text representation of SPMN is the output of $j - th$ Transformer in BERT, and visual and acoustic features are extracted by FaceT [37] and COVAREP [38], respectively. Immediately, the representations pass through the shared-private memory networks to obtain disentangled representations. The shared representations are X_{SL}, X_{SV}, X_{SA}, and the private representations are X_{PL}, X_{PV}, X_{PA}. Fig. 2

displays the information flow of shared-private memory networks. We then employ two adaptive fusion gate to extract text-related information from shared and private representations, and take the text-related information as inputs of $(j+1)$ Transformer encoder layer for pre-trained language representation model fine-tuning. Eventually, the first element $[CLS]$ of BERT is extracted to pass through fully-connected layers to make sentiment predictions.

3.2 BERT

As shown in Fig. 2, BERT contains M=12 Transformer encoder layers. SPMN integrates multimodal information into the j-th Transformer encoder layer and inputs it to the $(j+1)$-th Transformer encoder layer. Given a sentence $Z = [Z_1, Z_2, ...Z_N]$ of length N, $[CLS]$ is appended in front of the sentence as a marker of the predicted label. The representation of text is represented by $E = [E_1, E_2, ...E_N]$ through the input embedding layer that encodes each word with its embeddings, position embedding, and segmentation embedding. BERT takes E as input and generates a deep representation at each Transformer encoder layer. Suppose that the output of the j-th Transformer encoder layer is $X_L = [X_{l1}, X_{l2}, .., X_{lN}]$, and inputs the sequence of X_L together with $X_A = [X_{a1}, X_{a2}, ...X_{aN}]$ and $X_V = [X_{v1}, X_{v2}, ..., X_{vN}]$ are fed into the SPMN module to generate a new representation about multimodal information $\bar{X}_L$. Then $\bar{X}_L$ is fed into the $(j+1)$-th Transformer encoder layer. BERT integrates multimodal information via the SPMN module. Eventually, We pool the output of the last layer of the BERT. Specifically, the first elements $[CLS]$ of the BERT are extracted to pass through fully-connected layers to make sentiment predictions.

3.3 Memory Networks Construction

The memory networks $EncM$ employs a learnable matrix to preserve information for multimodal information exchange. Specifically, the learnable matrix contains multiple memory items $M \in R^{\kappa \times d}$, and each memory item is a vector $m_i \in R^{1 \times d}$, where κ and d are the number and dimension of memory items respectively. Following previous works [39], the memory items will have initialized values sampled from a uniform distribution. The memory items are initialized as follows:

$$g = \sqrt{\frac{2}{(1 + ns^2)}} \tag{1}$$

$$bound = g \times \sqrt{\frac{3}{\kappa}} \tag{2}$$

$$M \sim \mathcal{U}(-bound, bound) \tag{3}$$

where $\mathcal{U}$ and $bound$ are uniform distribution and uniform bound, respectively. g and ns denote the recommended gain value and negative slope in previous work [39], respectively.

Given the above memory items, we employ two independent fully connected layers to embed the memory items into pairs of keys and values respectively. We construct memory networks $EncM$ based on the above memory items and attention mechanism [18]. It is divided into two steps: memory query and memory response. The memory query and memory response are introduced in detail as follows. *Memory Query.* Given κ memory items, which is denoted by $M = \{m_1, m_2, ...m_\kappa\}$, The process of memory query can be formulated as:
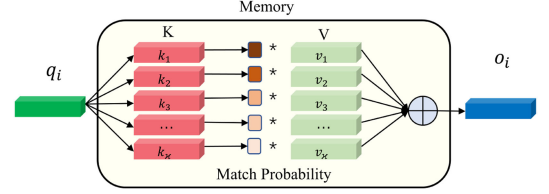

Fig. 3. Illustration of query and response from memory items. To query items in the memory, we compute the match probability w_i between query vector q_i and items $(m_1, m_2, ..., m_\kappa)$ and apply a weighted average of the items with the probabilities to obtain the response vectors.

$$Q = X_* \cdot W_q, * \in \{L, V, A\} \tag{4}$$

$$Q = \{q_1, q_2, ..., q_N\} \tag{5}$$

$$M = \{m_1, m_2, ...m_\kappa\} \tag{6}$$

$$K = M \cdot W_k = \{k_1, k_2, ..., k_\kappa\} \tag{7}$$

$$V = M \cdot W_v = \{v_1, v_2, ..., v_\kappa\} \tag{8}$$

where W_k and W_q are trainable weights. L, V, and A denote text, visual, and acoustic modalities, respectively. N denotes the sequence length of modality. We calculate the distances between the memory items and the query vectors as follows:

$$D_{ij} = \frac{q_i \cdot k_j}{\sqrt{d}} \tag{9}$$

where d represents the feature dimension. According to the above distances [35], we calculate the distances between query vectors and memory items, which are denoted by $D_i = \{D_{i1}, D_{i2} \cdots D_{i\kappa}\}$, where κ is the number of memory items. Then, the weights of the queried memory vector are calculated by:

$$w_{ij} = \frac{exp(D_{ij})}{\sum_{j=1}^{\kappa} exp(D_{ij})} \tag{10}$$

Memory Response. To employ all the memory items, we obtain all response vectors by:

$$o_j = \sum_{j=1}^{\kappa} w_{ij} v_j \tag{11}$$

where w_{ij} is the weights obtained from the memory query, v_j is the memory vector, and o_i is the queried memory vector (i.e.,, response vector). Fig. 3 shows the process of a single vector query from memory items and obtain the memory response. The outputs of memory network $EncM$ for query vectors are defined as:

$$O = EncM(X, M) \tag{12}$$

$$O = \{o_1, o_2, ..., o_N\} \tag{13}$$

where X and M denote query vectors and memory items, respectively. For simplicity, we adopt $EncM$ to represent the above memory query and memory response (i.e., memory networks). The N denotes the sequence length of modality.

3.4 Shared-Private Memory Networks

Let X_L, X_V, and X_A be three representations belonging to the text, visual, and acoustic modalities, respectively. The representation is split into two parts for each modality: the shared and private representations where the shared

TABLE 1
Predicted Results of SPMN on CMU-MOSI and CMU-MOSEI Datasets

Model	MOSI				MOSEI			
	Acc-2 ↑	F1-Score ↑	MAE ↓	CC ↑	Acc-2 ↑	F1-Score ↑	MAE ↓	CC ↑
TFN (G) [11]	-/80.8	-/80.7	0.901	0.698	-/82.5	-/82.1	0.593	0.700
LMF (G) [12]	-/82.4	-/82.4	0.917	0.695	-/82.0	-/82.1	0.623	0.677
MFN (G) [19]	77.4/-	77.3/-	0.965	0.632	76.0/-	76.0/-	-	-
RAVEN (G) [20]	78.0/-	76.6/-	0.915	0.691	79.1/-	79.5/-	0.614	0.662
MFM (G) [7]	-/81.7	-/81.6	0.877	0.706	-/84.4	-/84.3	0.568	0.717
MulT(Glove) [1]	81.5/84.1	80.6/83.9	0.861	0.711	-/82.5	-/82.3	0.580	0.703
MISA(B) [9]	81.8/83.4	81.7/83.6	0.783	0.761	83.6/85.5	83.8/85.3	0.555	0.756
MTAG(G) [14]	-/82.3	-/82.1	0.866	0.722	-	-	-	-
PMR(G) [10]	-/83.6	-/83.4	-	-	-/83.3	-/82.6	-	-
ICCN(B) [8]	-/83.07	-/83.02	0.862	0.714	-/84.18	-/84.15	0.565	0.713
Self-MM(B) [10]	84.0/86.0	84.4/85.9	0.713	0.798	82.8/85.2	82.5/85.3	0.530	0.765
M3SA(B) [41]	-/85.7	-/85.6	0.714	0.794	-/85.6	-/85.5	0.587	0.789
MMIM(B) [21]	84.14/86.06	84.00/85.98	0.700	0.800	82.24/85.97	82.66/85.94	**0.526**	0.722
BBFN(B) [22]	-/84.3	-/84.3	0.776	0.755	-/86.2	-/86.1	0.529	0.767
MAG(B) [15]*	84.37/85.19	84.68/85.04	0.730	0.794	84.79/85.28	84.72/85.17	0.594	0.784
MAG(X) [15]*	85.71/86.41	85.60/86.39	0.693	0.820	85.7/86.08	85.6/86.08	0.595	0.793
SPMN(B) (ours)	85.68/86.56	85.58/86.50	0.707	0.803	85.72/86.35	85.57/86.36	0.578	0.795
SPMN(X) (ours)	**87.04/88.02**	**86.88/87.94**	**0.667**	**0.830**	**86.41/86.77**	**86.39/86.73**	0.571	**0.800**

the numbers before '/' denote the average results of 5 runs, while the numbers after '/' denote the best results of 5 runs. Models with '*' are reproduced under the same conditions on the CMU-MOSEI. (G), (B), (X) indicate that the text embedding is extracted by Glove [40], BERT [16], and XLNET [17], respectively. '↑' indicates good performance for large values, and '↓' indicates good performance for small values.

representations contain the common information among the three modalities and the private representations contain the specific information for each modality. Therefore the representation of modality $X_*, * \in \{L, A, V\}$ can be divided into $\{X_{S*}, X_{P*}\}, * \in \{L, V, A\}$ by employing shared-private memory networks.

We build the shared-private memory networks based on memory networks $EncM$ in Section 3.3. For shared representations, we introduce shared memory items M_S. Similarly, for private representations, we introduce private memory items M_{PL}, M_{PV}, and M_{PA}. We take paired modality representations and memory items as the inputs of $EncM$, and obtain corresponding disentangled representations. Concretely, each modality queries from the shared memory items and obtains the shared representations, and queries from the private memory items to obtain private representations. Initially, there is only one representation for each modality (i.e., text (X_L), visual (X_V), acoustic (X_A)). The $EncM$ takes three paired modality representations and shared memory items (X_L, M_S), (X_V, M_S), and (X_A, M_S) as its inputs, and outputs the shared representations X_{SL}, X_{SV}, X_{SA}:

$$M_S = \{m_{s1}, m_{s2}, .., m_{sk}\} \quad (14)$$
$$X_{SL} = EncM(X_L, M_S) \quad (15)$$
$$X_{SV} = EncM(X_V, M_S) \quad (16)$$
$$X_{SA} = EncM(X_A, M_S) \quad (17)$$

where X_{SL}, X_{SV}, and X_{SA} denotes three shared representation. The M_S denotes shared memory items.

Similarly, the memory networks $EncM$ takes three paired modality representations and corresponding private memory items (X_L, M_{PL}), (X_V, M_{PV}), (X_V, M_{PV}) as inputs and outputs three private presentations.

$$M_{PL} = \{m_{pl1}, m_{pl2}, .., m_{plK}\} \quad (18)$$
$$M_{PV} = \{m_{pv1}, m_{pv2}, .., m_{pvK}\} \quad (19)$$
$$M_{PA} = \{m_{pa1}, m_{pa2}, .., m_{paK}\} \quad (20)$$
$$X_{PL} = EncM(X_L, M_{PL}) \quad (21)$$
$$X_{PV} = EncM(X_V, M_{PV}) \quad (22)$$
$$X_{PA} = EncM(X_A, M_{PA}) \quad (23)$$

where X_{PL}, X_{PV}, and X_{PA} denotes three private representations. The M_{PL}, M_{PV}, and M_{PA} are text, visual, and acoustic private memory items, respectively.

3.5 Adaptive Fusion Gate

SPMN integrates multimodal information in the process of BERT fine-tuning. Pre-trained language representation model does not have the components to directly accept heterogeneous multimodal information. Therefore, we exploit the adaptive fusion gate (AFG) to filter out irrelevant non-text information and screen out the most relevant non-text information to the text embedding. The architecture of the adaptive fusion gate is shown in Fig. 1. It takes shared and private features as inputs and outputs the dynamic filter results. The shared and private representations X_{SL}, X_{SV}, X_{SA} and X_{PL}, X_{PV}, X_{PA} are processed via the following adaptive fusion gate:

$$\alpha_{pv} = Sigmod([X_{PV}; X_{PL}] \cdot W_{PV} + b_{PV}) \quad (24)$$
$$\alpha_{pa} = Sigmod([X_{PA}; X_{PL}] \cdot W_{PA} + b_{PA}) \quad (25)$$
$$\alpha_{sv} = Sigmod([X_{SV}; X_{SL}] \cdot W_{SV} + b_{SV}) \quad (26)$$
$$\alpha_{sa} = Sigmod([X_{SA}; X_{SL}] \cdot W_{SA} + b_{SA}) \quad (27)$$
$$\hat{X}_{SL} = \alpha_{sv} \otimes X_{SV} + \alpha_{sa} \otimes X_{SA} + X_{SL} \quad (28)$$
$$\hat{X}_{PL} = \alpha_{pv} \otimes X_{PV} + \alpha_{pa} \otimes X_{PA} + X_{PL} \quad (29)$$
$$\bar{X}_L = X_L + \hat{X}_{SL} + \hat{X}_{PL} \quad (30)$$

TABLE 2
Distribution Statistics of the Datasets CMU-MOSEI
and CMU-MOSI

Dataset	Train	Valid	Test	All
CMU-MOSEI	16326	1871	4659	22856
CMU-MOSI	1284	229	686	2199

where X_{PL}, X_{PV}, X_{PA}, X_{SL}, X_{SV}, and X_{SA} are the shared and private representations. $\hat{X}_{SL}$ and $\hat{X}_{PL}$ represent the text-related information, where the visual and acoustic information irrelevant to the text are filtered out through the adaptive fusion gate. '$\otimes$' represents elements multiplication. '[;]' denotes concatenation in feature dimension. $Sigmod$ denotes activation function. $\bar{X}_L$ represents the input of $(j+1)$ Transformer encoder layer. W_{*L}, W_{*V}, W_{*A}, b_{*L}, b_{*V}, and b_{*A} are learnable parameters.

4 OPTIMIZATION OBJECTIVE

4.1 Difference Loss

The difference loss is also applied to shared and private representation and ensures both shared, and private memories capture different aspects of the inputs. The distance measure of two representations is implemented by utilizing a soft subspace orthogonality constraint [9], [42]. X_{S*} and X_{P*} denote matrices that are shared and private representations of the same modality in Section 3.4. The difference loss encourages orthogonality between the shared and the private representations of the same modality. We define the difference loss:

$$\mathcal{L}_{diff} = \sum_{*\in\{L,A,V\}} \left\| X_{S*}^T X_{P*} \right\|^2 \quad (31)$$

4.2 Supervised Loss

The task-specific loss estimates the quality of prediction during training. We use L2 loss as the fundamental optimization objective in Eq. 24, y and $\hat{y}$ denote ground truth and prediction, respectively.

$$\mathcal{L}_{task} = \frac{1}{N} \sum_{i=1}^{N} \left\| y_i - \hat{y}_i \right\|^2 \quad (32)$$

The overall loss of the model is defined to be a weighted sum of the difference loss and supervised loss. The weight α is a hyperparameter.

$$\mathcal{L} = \mathcal{L}_{task} + \alpha\mathcal{L}_{diff} \quad (33)$$

5 EXPERIMENTS

5.1 Dataset

We follow the prior works [1], [10], [15], and conduct experiments on two datasets: CMU-MOSEI [13] and CMU-MOSI [43]. CMU-MOSI is a multimodal sentiment analysis dataset composed of 2199 YouTube video clips. Each multimodal sample has a sentiment score distributed in [-3, 3], where 3 means strongly positive, and -3 means strongly negative. CMU-MOSEI is a dataset of movie reviews collected from YOUTUBE for sentiment analysis. It contains
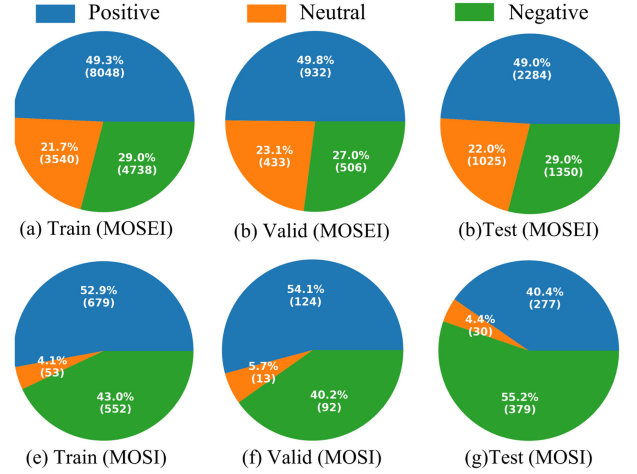

Fig. 4. The data distribution of each sentiment category in the training set, validation set, and test set.

22856 video clips. Table 2 shows the distribution of the dataset, and Fig. 4 shows the sentiment distribution of sentiment categories on CMU-MOSI and CMu-MOSEI. The following metrics are used to evaluate the performances of all models: Binary Classification Accuracy (Acc-2), F1-Score, Mean Absolute Error (MAE), and Correlation Coefficient (CC). In addition, binary classification accuracy (Acc-2) is calculated by converting the regression output into categorical values. A higher value means better performance for all the metrics except MAE. The above evaluation metrics are consistent with the previous work [1], [10], [15], [44].

5.2 Feature Extraction

For fair comparisons, we produce the standard machine-understandable low-level features. These features are employed by the CMU-MOSI and CMU-MOSEI benchmarks and utilized by the previous works [15]. Concretely, we process the text, visual, and acoustic features into typical tensors as introduced below. *Text Modality*. Many previous works utilized Glove [16] as text embedding. Following current state-of-the-art work [15], we adopt the pre-trained BERT [16] and XLNET [17] as feature extractors for text modality. However, the state-of-the-art works [10], [22] are based on BERT features. Thus, for fair comparisons, we provide results using both BERT and XLNET. For the pre-trained language representation model, we utilize the base pre-trained model, such as "BERT-base-uncased" and "XLNET-base-cased". The pre-trained language representation model includes 12 stacked Transformer encoder layers and an embedding encoder layer. Aligned with recent works [15], we utilize the output of a specific BERT encoder layer. Concretely, our model first uses output of BERT's embedding layer as the text representation. In addition, we also conduct experiments using each Transformer encoder layer of BERT as the text representation. *Visual Modality*. Following current state-of-the-art work [15], both CMU-MOSI and CMU-MOSEI adopt the FaceT[37] library to extract a set of basic emotions purely from static faces. Besides, CMU-MOSEI employs Facial Action Coding System (FACS) [45] and MultiComp OpenFace [46] to extract facial action units, facial landmarks, head pose, gaze

TABLE 3
The Hyperparameter Settings Adopted in Each Multimodal
Sentiment Analysis Benchmark

Setting	CMU-MOSEI	CMU-MOSI
Optimizer	Adam	Adam
Batch size	32	32
Learning rate	1e-4	1e-5
Epoch number	150	150
Alpha	0.5	1
Memory size	8	16
Dropout	0.5	0.5
NS (ns)	0	0

tracking, and HOG features. As a result, the final visual feature dimensions are 47 for CMU-MOSI and 35 for CMU-MOSEI.

Acoustic Modality. The acoustic features contain complex sentiment-relevant features, including fundamental frequency, 12 Mel-frequency cepstral coefficients, Voiced/Unvoiced segmenting features, etc. Following previous work [15], we utilized COVAREP [38] to extract the sentiment-relevant features. The feature dimensions are 74 for CMU-MOSI and CMU-MOSEI.

Modality Alignment. The extracted features in our experiments were word-aligned. Following previous work [15], we first perform a word-level forced alignment among text, visual and acoustic to determine when particular words appear in the visual and acoustic segment. Following many previous works [1], [15], [22], we adopted P2FA [47] to align visual and acoustic segments to each word. The tool automatically separates the co-occurring acoustic and visual frames into a group, and the frame features are averaged to a representation vector as a token. We repeat this operation for each word to calculate the visual and audio tokens.

5.3 Implementation Details

In our experiments, we use the same acoustic and visual features as the previous work [1]. The text feature is the j-th Transformer encoder layer of BERT. We adopt the Adam optimizer to train the network for 150 epochs. Table 3 displays the hyperparameters in detail.

5.4 Baselines

To verify the effectiveness of SPMN, we compare SPMN with the following state-of-the-art methods: *TFN (Tensor Fusion Network)* [11] is a tensor fusion method, which captures the interaction relationship of multiple tensor. *LMF (Low-rank Multimodal Fusion)* [12] solves the problem that the tensor dimension increases exponentially with the number of modalities. *MFN (Memory Fusion Network)* [13] is a multi-view modeling on time series data, using gated memory to control the flow of information. *MFM (Multimodal Factorization Model)* [7] is a discriminant model for joint optimization of multimodal data and label generation. *RAVEN (Recurrent Attended Variation Embedding Network)* [20] considers the fine-grained structure of non-verbal sub-word sequences and the dynamic transfer word representation based on non-verbal clues. *MULT (Multimodal Transformer)*

[1] introduces a Transformer-based crossmodal information fusion method. *ICCN (Interaction Canonical Correlation Network)* [8] considered the importance of text in multimodal sentiment analysis. *MAG (Multimodal Adaptation Gate)* [15] is to integrate multimodal information in the process of fine-tuning the Pre-trained model. *MISA (Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis)* [9] considers the similarity and difference among the the different modalities. *MATG (Modal-Temporal Attention Graph)* [14] is an interpretable graph neural network model that converts unaligned multimodal data into heterogeneous edges and nodes. *PMR (Progressive Modality Reinforcement)* [4] designs a message center to interact with unimodal information. *Self-MM (Self-Supervised Multi-task Multimodal sentiment analysis network)* [10] introduces a self-supervised method to learn modal specific information. *M3SA (Modulation Model for Multimodal Sentiment Analysis)* [41] identifies the contribution of modalities and reduce the impact of noisy information. *MMIM (MultiModal InfoMax)* [21] maximizes the Mutual Information between paired modalities. *BBFN (Bi-Bimodal Fusion Network)* [22] performs fusion and separation on pairwise modality representations.

5.5 Quantitative Analysis

5.5.1 Performance Comparison

To justify the effectiveness of our proposed SPMN model, we compared it with the several state-of-the-art baselines in the multimodal sentiment analysis task. As shown in Table 1, SPMN outperforms other methods in most cases. Note that we directly quoted the results of these baselines from their original papers, except for MAG [15]. Following previous works [15], we also utilized BERT [16] and XLNET [17] as the representation of text modality for fair comparisons. The comparison results are summarized in Table 1. By analyzing this table, we gained the following observations:

- Among modality-invariant and modality-specific methods, BBFN [22] surpasses MISA [9] by a large the margin on two datasets. MISA employs different losses to constrain the multimodal representation, BBFN employs the adversarial training approach. Specifically, it builds a compound gate and a retain gate for information flow control. The fact indicates that elaborately modeling unimodal representation is extremely essential for multimodal sentiment analysis.
- The pre-trained language representation model has achieved significant improvements in the downstream NLP tasks. Regarding how to use the pre-trained language representation model, MAG [15] outperforms Self-MM [10] on all criteria of the two datasets. Because it could capture high-order correspondences via multiple BERT encoder layers, verifying the importance of integrating acoustic and visual information on a specific BERT encoder layer.
- Our proposed model SPMN outperforms the compared baselines regarding Acc-2 on CMU-MOSI and CMU-MOSEI. Compared with MAG [15], our approach achieves improvement with nearly 1.31% and 0.93% Acc-2 average gain on the CMU-MOSI and CMU-MOSEI test sets, respectively. Compared

TABLE 4
Performance Comparison on CMU-MOSI and CMU-MOSEI With Different Text Embeddings

Variant Model	MOSI				MOSEI			
	Acc-2↑	F1-Score↑	MAE↓	CC↑	Acc-2↑	F1-Score↑	MAE↓	CC↑
BERT[16]	84.27	84.14	0.784	0.750	84.12	84.12	0.625	0.760
RoBerta [24]	84.12	84.16	0.783	0.770	85.03	84.81	0.641	0.781
DeBerta [25]	83.66	83.60	0.773	0.768	85.28	85.07	0.596	0.780
XLNET [17]	85.95	85.92	0.703	0.816	85.69	85.62	0.600	0.786
Electra [26]	87.33	87.31	0.713	0.819	86.05	86.03	0.562	0.811
SPMN(B)	86.56	86.50	0.707	0.803	86.35	86.36	0.578	0.795
SPMN (R)	85.34	85.33	0.743	0.784	86.46	86.43	0.578	0.796
SPMN (D)	85.65	85.56	0.708	0.804	86.71	86.55	0.569	0.799
SPMN (X)	88.02	87.94	0.667	0.830	86.77	86.73	0.571	0.800
SPMN (E)	**89.97**	**89.93**	**0.600**	**0.866**	**87.35**	**87.34**	**0.550**	**0.821**

The best results are highlighted in bold. (B), (R), (D), (X), (E) indicate that the text embedding are extracted by BERT, Roberta, Deberta, XLNET, and Electra, respectively.

to previous work [15], [21], [22], Experiments demonstrated that our model could achieve the best performance on both CMU-MOSI and CMU-MOSEI datasets. Likewise, for multimodal sentiment analysis, our model produces the best results over previous methods on most metrics. The improvement indicates the feasibility and importance of dynamically exploring modality interaction patterns.

- The results of our model SPMN (BERT) and SPMN (XLNET) are competitive to some state-of-the-art methods, especially the prior ensemble models (e.g., MAG [15] and BBFN [22]). This further demonstrates the effectiveness and robustness of our proposed model and indicates the remarkable ability of our shared and private modules.

In addition, we also conducted the significance test [48] over the regression prediction results between SPMN (B) and the state-of-the-art model MAG (B). To be specific, on CMU-MOSI, the p-values are 4.0E-5. It can be seen that these p-values are observably smaller than 0.05, indicating the statistically significant advantage of our model SPMN.

5.5.2 Analysis of Generalization Ability

We also conduct experiments to verify that our proposed SPMN is generalized to be applied. To gain deep insights into our proposed shared-private memory networks, we illustrated several results with different pre-trained language representation models. Concretely, we employed different pre-trained language representation models as the BERT backbone to integrate multimodal information. Such as BERT [16], XLNET [17], RoBerta [24], DeBerta [25], and Electra [26]. The results obtained from CMU-MOSI and CMU-MOSEI are displayed in Table 4, respectively. We gained the following observations. 1) Our proposed model can efficiently integrate multimodal information into the BERT backbone. On BERT, XLNET, RoBerta, DeBerta, and Electra, SPMN received absolute improvements of 2.29%, 2.22%, 1.99%, 2.07%, and 2.64% in the Acc2 on CMU-MOSI dataset, respectively. And 2) our proposed model is universal, which can integrate multimodal information for different pre-trained language representation models.

To explore the impact of applying SPMN, we apply SPMN at different encoder layers of the pre-trained language

representation model [15]. Specifically, we first integrate multimodal information into the BERT embedding layer by utilizing SPMN. Subsequently, we apply the SPMN to the pre-trained language representation model's layer $1 <= j <= 12$. From Fig. 5, we observed that our proposed SPMN integrates multimodal information at earlier layers of the BERT backbone and achieves better performance. The facts demonstrate that earlier layers are more suitable for applying SPMN. Initially, visual and acoustic modalities are low-level semantic information, and text modality contains different levels of semantic [49] at different BERT encoder layers. Regarding the syntactic and semantic structure of text modality features, the higher encoder layers of the BERT model abstract and higher-level information. Therefore, integrating low-level visual and acoustic features with low-level textual features (i.e., the initial embedding layer of BERT) is more suitable for the application of SPMN.

5.5.3 Memory Analysis

To further explore the impact of the memory networks, we conducted experiments by increasing memory size (i.e., the number of memory vectors) from 4 to 64. The results are shown in Fig. 6. From the comparison results, we could find that enlarging the number of memory vectors in an appropriate range can improve the multimodal sentiment analysis
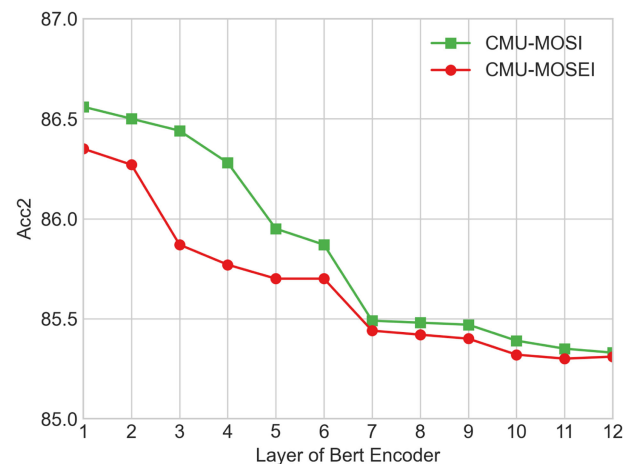

Fig. 5. Results comparison on CMU-MOSI and CMU-MOSEI applying SPMN at different BERT encoder layers.
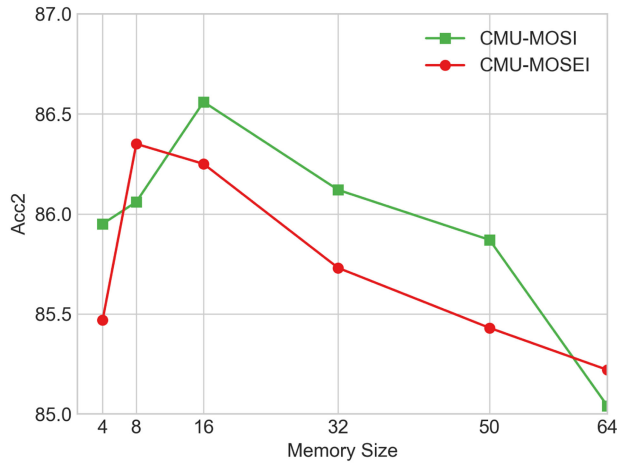
Fig. 6. The accuracy and the number of parameters of SPMN against the memory size.

TABLE 6
The Ablation Study on CMU-MOSI and CMU-MOSEI to Justify the Effect of Different Fusion Methods

Fusion	MOSI		MOSEI	
	Acc-2	F1-Score	Acc-2	F1-Score
Fc	86.28	86.24	86.08	86.06
Add	86.38	86.34	86.11	86.10
Attention	86.41	85.36	86.26	86.21
AFG	**86.56**	**86.50**	**86.35**	**86.36**

'Fc' denotes fully connected layer. 'Attention' denotes self-attention [18]. 'Add' means to add multimodal representations. 'AFG' denotes our proposed adaptive fusion gate.

performance, enhancing the model's representation ability. However, the performance begins to drop when memory items are greater than 16. The reason may be that they can not be fully updated, so they do not help the multimodal sentiment analysis other than being played as noise, limiting the model optimization and further hindering the sentiment prediction. From Fig. 6, we could see that enlarging memory size within the threshold improves performance, and similar patterns were observed in CMU-MOSI and CMU-MOSEI.

5.5.4 Ablation Study

To gain insights into our three modules, we conducted ablation studies incrementally. To be more specific, we compared the model SPMN with the following variants: 1) w/o different loss, removing the different loss; 2) w/o private module, eliminating the private memory and the different loss; 3) w/o shared module, without the shared memory and the different loss. As reported in Table 5, compared with our model SPMN, the performance of the w/o shared module degrades dramatically. Notably, it drops absolutely by 0.92% and 1.02% on Acc2 for SPMN on datasets CMU-MOSI and CMU-MOSEI, respectively. Besides, the performance drop of the w/o private memory can be observed, indicating that it is essential to consider the shared representation from shared memory to enhance the representations of multiple modalities. The w/o shared module and w/o private module demonstrate the vital importance of shared and private memories as they can learn two distinct

representations. Moreover, our model achieves better results than w/o different losses, revealing that the different losses can enhance the representations of shared and private memories and boost the model performance. In general, our proposed SPMN vastly exceeds all variants on multimodal sentiment analysis, verifying the effectiveness and complementarity of the three modules.

Previous work [1], [4], [22] mainly employed sophisticated approaches to model multimodal fusion networks sufficiently to obtain excellent results. Unlike these methods, our proposed model can achieve state-of-the-art performance by utilizing simple fusion strategies. From Table 6, even with simple and direct fusion methods like fully connected layers and self-attention, SPMN outperforms all baselines in most indicators. The comparison results show that shared and private memories are influential and satisfactory generalization abilities.

5.6 Qualitative Results

5.6.1 Shared-Private Representations Visualization

Apart from achieving superior performance, the critical advantage of SPMN over other methods is that its shared and private memory networks can assign memory items for different representations of different modalities. To this end, we visualized the shared and private representations learned by shared-private memory networks. Specifically, we first obtained the shared and private for each modality and then used t-SNE [50] to map the shared and private representations into the two-dimensional space. Afterward, we clustered these 2D representations into different groups by shared and private categories, where each group was marked in one color.

Fig. 7 (left) displays the original data distribution. From Fig. 7 (middle), the blue points (text modality), the orange

TABLE 5
The Ablation Study on CMU-MOSI and CMU-MOSEI to Investigate the Effect of Different Components of the SPMN

Variant Model	MOSI				MOSEI			
	Acc-2↑	F1-Score↑	MAE↓	CC↑	Acc-2↑	F1-Score↑	MAE↓	CC↑
BERT	84.27	84.14	0.784	0.750	84.12	84.12	0.625	0.760
SPMN (w/o diff loss)	86.11	86.10	0.714	0.800	86.16	86.13	0.580	0.793
SPMN (w/o private module)	85.39	85.27	0.741	0.787	85.41	85.35	0.588	0.786
SPMN (w/o shared module)	85.64	85.60	0.733	0.789	85.33	85.18	0.591	0.784
SPMN(B) (ours)	**86.56**	**86.50**	**0.707**	**0.803**	**86.35**	**86.36**	**0.578**	**0.795**

The best results are highlighted in bold.

2898 IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 14, NO. 4, OCTOBER-DECEMBER 2023
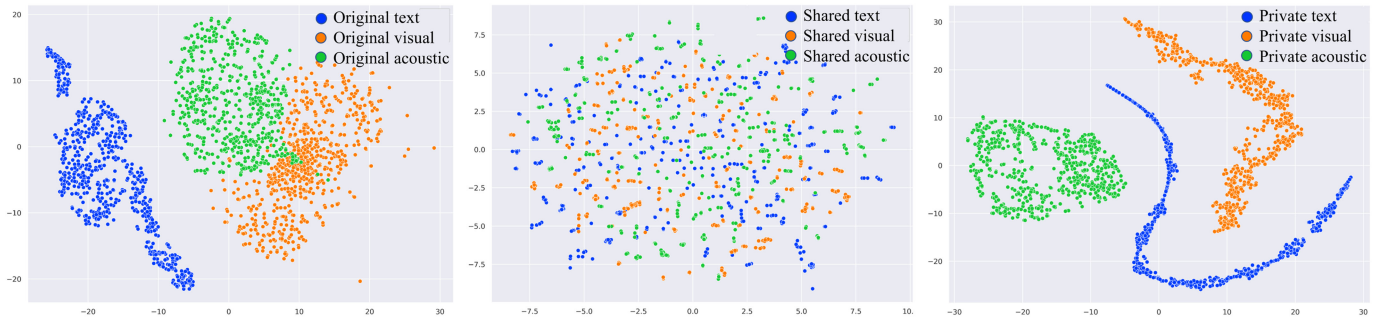

Fig. 7. Visualization of the shared and private representations of three modalities in the testing set of MOSI datasets using t-SNE [50] projections. Observations on MOSEI are also similar. The left subfigure represents the original multimodal representations, the middle subfigure represents the multimodal shared representations, and the right subfigure represents the multimodal private representations.

Fig. 8. Visualization of the shared and private representations of the same modality in the testing set of MOSI datasets using t-SNE [50] projections. Observations on MOSEI are also similar. The left subfigure represents the text shared and private representations, the middle subfigure represents the visual shared and private representations, and the right subfigure represents the acoustic shared and private representations.

points (visual modality), and the green ones (acoustic modality) are uniformly distributed in the 2D plane. Different modalities pass through shared memory networks to obtain the shared representations. Fig. 7 (right) shows the private representations. there is a large margin between the blue points (text modality), the orange points (visual modality), and the green ones (acoustic modality), as they are clearly distinct modalities. From Fig. 8, we could see that two representations of the same modality can be well distinguished according to the distribution of learned shared and private representations. The shared-private memory networks can well learn the shared and private representations of different modalities. In addition, as shown in Fig. 9, we also used t-SNE to visualize the shared multimodal representations and the private multimodal representations distribution along the annotations, respectively. These results demonstrate that 1) our model can intelligently understand

shared and private representations of the same modality by utilizing shared-private memory networks. Therefore the distribution of learned representations is consistent with that of semantics to some extent. And 2) our proposed shared-private memory networks are capable of perceiving the share and private parts of multimodal data.

5.6.2 Case Study

To qualitatively validate the effectiveness of SPMN, we displayed several typical examples of CMU-MOSI and CMU-MOSEI in Fig. 10, respectively. Based on these multimodal sentiment analysis results (i.e., A, B, C), we could see that our model could comprehend positive, negative, and neutral sentiments accurately. Meanwhile, it is robust for different sentiment polarity, mainly attributed to our proposed model's decoupled representation. Specially, we compared the difference between our methods and MAG for the prediction. SPMN(BERT) provides closer scores to the ground truths than BERT and MAG-BERT, owing to the shared and private memories. The cases indicate that SPMN can effectively integrate non-verbal modalities with verbal modality and is better than MAG. In addition, we also analyzed the wrong examples to indicate the shortcomings of the model. As shown in Fig. 10, Our model achieves incorrect classification results on cases D and E. We analyzed the content of these cases. The visual and acoustic modalities of the case show neutral emotion, while the text modality tends to describe the content of the movie rather than the author's sentiment toward the movie. Therefore, the bias of the text modality leads to wrong classification results.
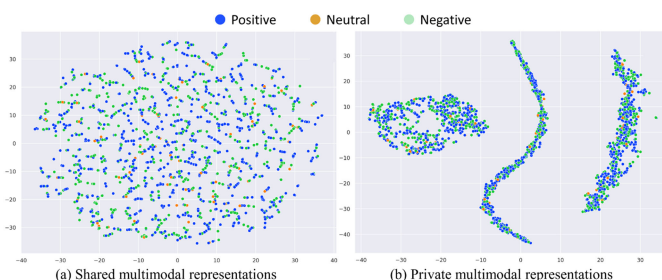

Fig. 9. Visualization of multimodal representations (i.e., text, visual, and acoustic modalities.) distribution along the annotations. The labels are positive, negative and neutral sentiments respectively. (a): Visualization of shared representations. (b) Visualization of private representations
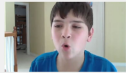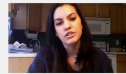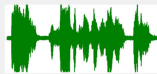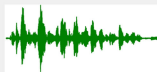
Case	Spoken words	Visual behaviors	Acoustic features	Ground Truth	**SPMN**	MAG	BERT
A	The action is fucking awesome.			+2.79	**+2.79**	+2.66	+2.88
B	When I saw it was just totally boring.			-2.49	**-2.40**	-2.15	+2.07
C	The first time she goes in it looks like a Japanese temple.			+0.02	**+0.00**	+0.10	+0.16
D	I was saying if she not gonna marry him I will.			+0.80	**-0.67**	-1.00	-1.21
E	The first two were great and they brought in the kids and it made it more kiddy friendly movie.			-0.80	**+2.00**	+1.91	+2.58

Fig. 10. Input of there modalities and prediction with different model on CMU-MOSI in our case study. Positive, negative and neutral sentiments predicted on CMU-MOSI.

6 CONCLUSION

In this work, we propose a novel perspective for modeling the representations of multimodal data that are shared and private memory networks. Our goal is to model efficient shared and private representations and fuse the representations for pre-trained language representation model fine-tuning, then obtain considerable performance in multimodal sentiment analysis. Experimental results demonstrate that SPMN achieves state-of-the-art performance on CMU-MOSEI and CMU-MOSI datasets.

REFERENCES

[1] 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," in *Proc. Conf.. Assoc. Comput. Linguistics*, 2019, Art. no. 6558.

[2] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Trans. Affect. Comput.*, vol. 13, no. 01, pp. 320–334, First Quarter 2022.

[3] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 4477–4481.

[4] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2554–2562.

[5] S. Katada, S. Okada, and K. Komatani, "Effects of physiological signals in different types of multimodal sentiment estimation," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3155604.

[6] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3171091.

[7] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, arXiv: 1806.06176.

[8] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.

[9] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[10] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10 790–10 797.

[11] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[12] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[13] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018a, pp. 2236–2246.

[14] J. Yang et al., "MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2021, pp. 1009–1021.

[15] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, Art. no. 2359.

[16] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[20] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.

[21] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[22] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L. P. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proc. Int. Conf. Multimodal Interaction*, 2021, pp. 6–15.

[23] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2978–2988.

[24] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv: 1907.11692*.

[25] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced bert with disentangled attention," 2020, *arXiv: 2006.03654*.

[26] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv: 2003.10555*.

[27] J. Weston, S. Chopra, and A. Bordes, "Memory networks," 2014, *arXiv:1410.3916*.

[28] S. Sukhbaatar et al., "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.

[29] A. Miller, A. Fisch, J. Dodge, A.-H. Karimi, A. Bordes, and J. Weston, "Key-value memory networks for directly reading documents," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 1400–1409.

[30] A. Kumar et al., "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1378–1387.

[31] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 372–14 381.

[32] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, "Memory-attended recurrent network for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8347–8356.

[33] M. Zhu, P. Pan, W. Chen, and Y. Yang, "DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5802–5810.

[34] L. Zhu and Y. Yang, "Inflated episodic memory with region self-attention for long-tailed visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4344–4353.

[35] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5904–5914.

[36] M. Lee and V. Pavlovic, "Private-shared disentangled multimodal vae for learning of latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1692–1700.

[37] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*. Berlin, Germany: Springer, 2005, pp. 247–275.

[38] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep–A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[40] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[41] Y. Zeng, S. Mai, and H. Hu, "Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis," in *Proc. Findings Assoc. for Comput. Linguistics Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 1262–1274.

[42] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 343–351.

[43] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multi-modal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.

[44] X. Zhao, Y. Chen, W. Li, L. Gao, and B. Tang, "MAG+: An extended multimodal adaptation gate for multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4753–4757.

[45] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1125.

[46] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[47] J. Yuan et al., "Speaker identification on the scotus corpus," *J. Acoustical Soc. Amer.*, vol. 123, no. 5, 2008, Art. no. 3878.

[48] L. Qu, M. Liu, J. Wu, Z. Gao, and L. Nie, "Dynamic modality interaction modeling for image-text retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1104–1113.

[49] E. Reif et al., "Visualizing and measuring the geometry of BERT," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8594–8603.

[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Xianbing Zhao received the bachelor's degree from the Harbin Institute of Technology (Weihai), in 2016. Currently working toward the second year of my PhD degree with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. His research interests include multimodal sentiment analysis and medical artificial intelligence.

Yixin Chen received the BE degree in software engineering from the School of Computer Science, Nanjing Normal University, Nanjing, China, in 2020. Currently working toward the second year of master's degree with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China. Her research interests include multimodal NLP and radiology report generation.

Sicen Liu is currently working toward the PhD degree with the Harbin Institute of Technology, Shenzhen under the supervision of Prof. Xiaolong Wang in the field of medical informatics. Her current research interests include sequence learning, medical informatics, and artificial neural networks.

Buzhou Tang (Member, IEEE) received the PhD degree from the Harbin Institute of Technology, China, in 2011. He is currently a professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen. China. His research interests cover artificial intelligence, natural language processing, medical informatics and multimodal information processing.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.