Graph-based Fine-grained Multimodal Attention Mechanism for Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal sentiment analysis is a popular research area in natural language processing. Mainstream multimodal learning models barely consider that the visual and acoustic behaviors often have a much higher temporal frequency than words. Therefore, these models lack the representation capability to accurately model multimodal interactions. In this paper, we propose an attachment called Graph-based Fine-grained Multimodal Attention Mechanism (GFMAM), which can utilize the multimodal information from different subspaces to achieve accurate multimodal interactions. Firstly, the attachment further splits the information of every modality into multiple subspaces. Then, the fine-grained multimodal information from different subspaces is converted into multimodal interaction graphs dominant by the language modality. The multimodal interaction graph can capture significant interactions among multiple modalities at the subspace level. Finally, the information of nonverbal modalities is additionally added to compensate for the loss of continuity caused by the splitting operation. Embedding GFMAM into BERT, we propose a new model called GFMAM-BERT that can directly accept nonverbal modalities in addition to language modality. We conducted experiments on both publicly available multimodal sentiment analysis datasets CMU-MOSI and CMU-MOSEI. The experiment results demonstrate that GFMAM-BERT exceeds the state-of-the-art models. Moreover, the proposed model outperforms humans on most metrics on the CMU-MOSI dataset.

1 Introduction

People sharing their opinions, stories, and movie reviews on video sites like YouTube often involve the information of multiple modalities (language, visual, and acoustic). Since language may be misleading, a model relying solely on language information is insufficient to determine the speaker's affective state and correctly convey views and options (Williams et al., 2018). Therefore, multimodal sentiment analysis can provide better performance than the methods using only language modality, and it has received increasing attention. The central challenge of multimodal sentiment analysis is to model the *inter-modality* dynamics since the interactions among language, visual, and acoustic modalities can change the perception of the expressed sentiment (Zadeh et al., 2017).

To learn the relationships among modalities, many previous works summarize the information among modalities using simple averaging strategies (Sun et al., 2020; Hazarika et al., 2020; Yu et al., 2021). However, the visual and acoustic behaviors often have a much higher temporal frequency than language, leading to a sequence of accompanying visual and acoustic "subword" units for each uttered word (Wang et al., 2019b). Hence, the information of multimodalities requires fine-grained analysis. The previous works using the simple average strategies have not considered the utilization of the information from multiple subspaces to construct multimodal interactions. Although the simple average strategies may help to model global characteristics, it lacks its representational capacity to accurately model the structure of multimodal interactions at the subspace level. This motivated us to design a model that accurately captures the significant multimodal interactions from different subspaces.

We propose an attachment called GFMAM that can integrate fine-grained multimodal information from different subspaces. The attachment splits multimodal information into small granularities to obtain multiple feature subspaces. Then the fine-grained information is converted into multimodal interaction graphs to produce different sets of attention weights for different feature subspaces. In the graph, the fine-grained multimodal information and potential relationships between different modalities are represented as nodes and edges, respectively. The nodes of the language information are the core of the star-like graph since language modality contains more practical information than nonverbal modalities (Mai et al., 2019; Sun et al., 2020; Mai et al., 2021). Each fine-grained language node as a reference is connected to the nodes of the two other modalities (acoustic and visual modalities) to construct tri-modal interactions. The rich interaction between nodes in the graph neural network can help to capture significant interactions between different modalities. Moreover, since the splitting operation breaks the potential continuity within original information, we choose to compensate for the continuity within the nonverbal modalities with a residual approach. The multimodal sequence data is fed to GFMAM to obtain compact multimodal representations for shifting the word position (Wang et al., 2019b) in the semantic space.

We embed the attachment GFMAM into BERT, which can only process language modality, to give BERT the ability to accept and process nonverbal modalities directly. We evaluate GFMAM-BERT on two popular benchmark datasets of CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018d). The experiments show that the proposed method can produce better performance than the state-of-the-art methods, even outperform the exhibited humans on most metrics.

The contributions of this paper are therefore summarized as:

- We design a new multimodal interaction graph dominated by language modality. The graph is capable of capturing significant interactions between multiple modalities at the subspace level.

- We propose a Graph-based Fine-grained Multimodal Attention Mechanism (GFMAM) attachment achieving fine-grained multimodal information integration with the help of multimodal interaction graphs. Then, this attachment is successfully embedded into a large pre-trained model for the sentiment analysis.

- The proposed model outperforms the state-of-the-art methods. Furthermore, the results demonstrate that the proposed model surpasses the performance of humans for both binary classification and regression tasks on CMU-MOSI.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis is a popular research area in the field of NLP (Zadeh et al., 2018b). The expressive power of the single language modality is limited by the ambiguity of the language (Williams et al., 2018). The ambiguity usually appears in scenarios including the use of slang and sarcasm. To overcome the limitation of the single language modality, the additional information from multiple modalities can be a significant complement. The works of multimodal sentiment analysis can be divided into two categories according to whether the language modality is dominant.

Some methods consider the contribution of each modality is equal for sentiment analysis. Zadeh et al. (2017) created a multidimensional tensor by 3-fold Cartesian to capture uni-modal, bi-modal, and tri-modal interactions across three modalities. The LMF model proposed by Liu et al. (2018) decomposes the weights into low-rank factors, thus reducing the number of parameters in the model. This decomposition can be performed efficiently by using a parallel decomposition of the low-rank weight tensor and the input tensor to compute tensor-based fusion. Hazarika et al. (2020) used BERT to extract the feature information of the language modality and utilized two LSTMs (Hochreiter and Schmidhuber, 1997) to extract the acoustic and visual modality features. Each extracted modality feature is projected into two different spaces (modality-invariant and modality-specific). Then the information obtained from these projections is concatenated together for the sentiment analysis. However, in multimodal sentiment analysis or emotion recognition tasks, textual features usually outperform non-textual features (Sun et al., 2020; Mai et al., 2021; Sun et al., 2021). Therefore, the performance of these approaches is limited by the non-dominant role of the language modality.

The language modality is dominant in some methods. Wang et al. (2019b) modeled multimodal human language by shifting interactive word representations based on the text-video and text-audio interactions. The work done by Rahman et al. (2020) is an improvement of pre-trained models. Their proposed Multimodal Adaptation Gate (MAG) can change the word representations using both text-video and text-audio interactions. Then the new word representations are fed to large pre-trained transformers. Sun et al. (2020) constructed two outer product matrices ($T \otimes V$ and $T \otimes A$) to represent the text-video and text-audio interactions. The outer product matrices are then fed into a Canonical Correlation Analysis (CCA) network whose output is used for prediction. However, their cross-modal interaction is a bi-modal operation that only accounts for two modalities' input at a time. Thus, the proposed GFMAM enables tri-modal interactions in the fine-grained manner. And this interaction establishes the dominance of text features in multimodal sentiment analysis at the same time.

2.2 Graph Neural Networks

Graphs, a non-Euclidean data structure, have a great expressive power to model a set of objects (nodes) and their relationships (edges). Deep learning methods have succeeded in feature extraction of Euclidean data (e.g., images, text, and video). However, the traditional deep learning methods cannot effectively extract the features from the non-Euclidean data (Wu et al., 2020). Graph Neural Networks (GNNs) (Gori et al., 2005; Scarselli et al., 2008) started to try to extend deep neural networks to process the graph-structured data. After that, heterogeneous GNN methods (Wang et al., 2019a; Wei et al., 2019) are further proposed. Nodes in heterogeneous graphs represent different entities.

Recently, a few works have attempted to bring multimodal sequence data into graphs as a way to capture significant interactions among multi-modalities. Mai et al. (2020) employed graph convolutional networks for each modality to learn intra-modal dynamics. But it does not explicitly deal with cross-modality information. Wu et al. (2021) modeled multimodal sequence information with the graph-based neural model and capsule network. However, their cross-modal interaction is a bi-modal operation that only accounts for two modalities' input at a time. Based on these works, the natural dependency that exists between different modalities can be introduced to the graph structure. The most relevant work of this paper is the work proposed by Vaswani et al. (2017). They proposed Graph Attention Networks (GATs) using an attention mechanism to determine the weights of node neighborhoods when aggregating feature information. All multimodal information at the subspace level is converted into the heterogeneous nodes of the multimodal interaction graphs. Then the attention mechanism is used to find the attention coefficients between neighboring nodes to establish the dependencies between modalities.

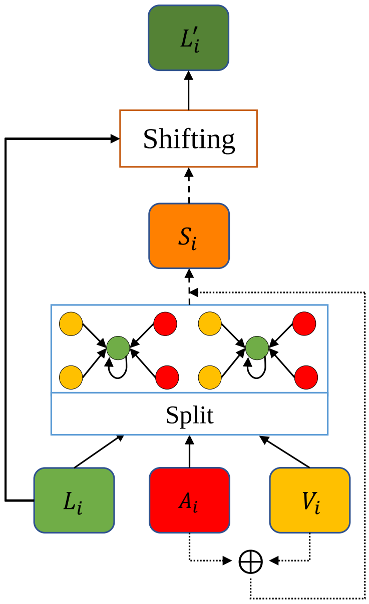

Figure 1: Overview of GFMAM attachment. Using language, visual and acoustic modalities of information as input, a representation of integrated multimodal information is obtained to shift the position of word in the semantic space. $\oplus$ denotes element-wise sum.

3 Graph-based Fine-grained Multimodal Attention Mechanism (GFMAM)

This section first introduces the splitting process of the multimodal information. Then, the construction of the multimodal interaction graph using fine-grained data is described in detail. Finally, we describe the complement of the modal internal continuity. The complete flow of the Graph-based Fine-grained Multimodal Attention Mechanism (GFMAM) is shown in Figure 1.

3.1 Fine-grained Multimodal Information

The GFMAM attachment accepts input from the three modalities language, visual, and acoustic modalities. The combination of these three modalities can reflect the emotional state of the speaker. Because the visual and acoustic behaviors often have a much higher temporal frequency than words. To capture the subtle variations among modalities, we separately split the information of three modalities into multiple subspaces.

We denote the multimodal information corresponding to the i-th word by a triple (L_i, V_i, A_i), where L denotes language features, V denotes visual features, and A denotes acoustic features. The language features are obtained by the embedding

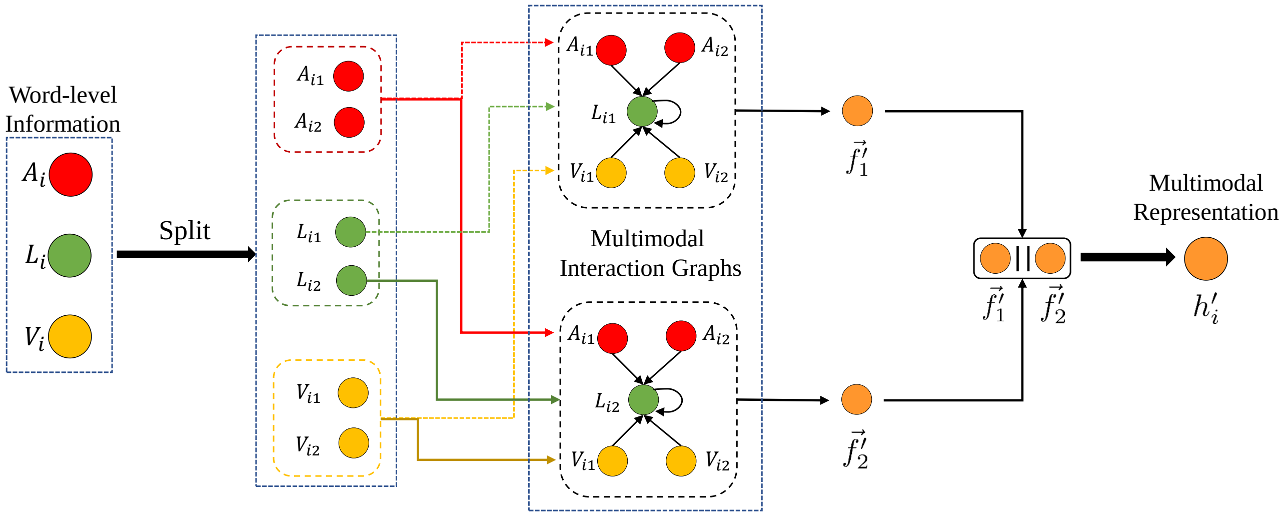

Figure 2: Examples of the multimodal interaction graph. Each modality information is split by $N = 2$. Then, the split results are integrated using the multimodal interaction graph. The final orange node h_i' denotes a new multimodal representation obtained after the operation. $||$ denotes concatenation.

layer. We use two subnets to extract visual and acoustic features, respectively. These extraction operations ensure that the features of the three modalities are in the same dimension. After extraction, a splitting operation is performed on the features of the three modalities. The multimodal features at the word level will be sequentially split into N parts. The fine-grained features of language modality are denoted as: $L_i = \{\overrightarrow{L_{i1}}, \overrightarrow{L_{i2}}, \cdots, \overrightarrow{L_{iN}}\}$, where N represents the number of split nodes. Similarly, the split results of visual and acoustic modalities can be denoted as: $V_i = \{\overrightarrow{V_{i1}}, \overrightarrow{V_{i2}}, \cdots, \overrightarrow{V_{iN}}\}$ and $A_i = \{\overrightarrow{A_{i1}}, \overrightarrow{A_{i2}}, \cdots, \overrightarrow{A_{iN}}\}$. The GFMAM accepts $h_i = \{\overrightarrow{L_{i1}}, \cdots, \overrightarrow{L_{iN}}, \overrightarrow{V_{i1}}, \cdots, \overrightarrow{V_{iN}}, \overrightarrow{A_{i1}}, \cdots, \overrightarrow{A_{iN}}\}$ as an input. To facilitate the representation, we use $h_i = \{\vec{f_1}, \vec{f_2}, \cdots, \vec{f_M}\}$ instead of the above equation, where $M = 3 * N$. The range $[1, N]$ indicates language modality information, $[(N+1), 2N]$ indicates visual modality information, and $[(2N+1), 3N]$ indicates acoustic modality information. We use the adjacency matrix in the implementation to control the connection relationship between nodes.

3.2 Multimodal Interaction Graph

After splitting the multimodal information into a smaller granularity, we convert the feature at the subspace level into the multimodal interaction graph. In this graph, the language modality can pay attention to the nonverbal modality features of its neighborhood. Then, the dependencies, which are the attention coefficient between nonverbal modalities and language modalities, are computed using the attention mechanism.

The importance of node j to node i is represented by e_{ij}, and a weight matrix $W \in R^{d\times d}$ is multiplied with each node to enhance the representation of nodes. The node-to-node attention coefficient is calculated using the function $a(\cdot) : R^d \times R^d \longrightarrow R^d$.

$$e_{ij} = a(W\vec{f_i}, W\vec{f_j}), \tag{1}$$

where the range of values of i is $i \in [1, N]$. We only compute e_{ij} for node $j \in \mathcal{X}_i$, where $\mathcal{X}_i$ is some neighborhood of node i in the graph (including i). To make coefficients easily comparable across different nodes, we normalize them across all choices of j using the softmax function:

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k\in\mathcal{X}_i} exp(e_{ik})} \tag{2}$$

Since a is a single-layer feedforward neural network, we add a nonlinear activation function $R(\cdot)$, which can be expressed as:

$$\alpha_{ij} = \frac{exp(R(a^T[W\vec{f_i}||W\vec{f_j}]))}{\sum_{k\in\mathcal{X}_i} exp(R(a^T[W\vec{f_i}||W\vec{f_k}]))} \tag{3}$$

where $.^T$ represents transposition and $||$ is the concatenation operation. The attention coefficients among nodes are then used to update each node that represents the language modality information. The i-th language modality node can be represented as:

$$\vec{f_i'} = R(\sum_{j\in\mathcal{X}_i} \alpha_{ij} W\vec{f_j}) \tag{4}$$

All nodes are then concatenated to obtain a compact multimodal representation.

$$h_i' = \{\vec{f_1'}||\vec{f_2'}||\cdots||\vec{f_N'}\} \tag{5}$$

The example as shown in Figure 2 demonstrates this operation for us.

3.3 Compensation of the Continuity within Modalities

The fine-grained multimodalities feature information from different subspaces can help to effectively capture the subtle variations. However, nonverbal modalities are usually presented continuously. Splitting operation may destroy the potential continuity within the nonverbal modalities. Therefore, we compensate for the continuity within the visual and acoustic modalities by a scaling factor $\beta(\beta < 1)$ to decrease the influence on the attention mechanism operation.

$$S_i = h_i' + \beta(V_i \oplus A_i) \tag{6}$$

where $\oplus$ denotes element-wise sum and β is a hyper-parameter selected through the cross-validation process. The obtained new multimodal information S_i is then used to shift the position of the word in the semantic space.

$$L_i' = L_i + S_i \tag{7}$$

4 GFMAM-BERT

GFMAM-BERT is GFMAM embedded between the embedding layer and the transformer layers of the BERT network. Recently, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), Transformer-based (Vaswani et al., 2017) contextual word representations, has shown excellent performance in multiple disciplines within NLP (Rahman et al., 2020). Therefore, BERT is chosen as the basis for sentiment analysis tasks in our work.

Figure 3 clearly shows the exact location of the GFMAM attachment embedded in the BERT. There are no changes to the BERT structure except for the attachment of GFMAM. The input to BERT is the original words in the language modality. A special token ($[CLS]$) is added in front of each sentence of the input, which is processed by the transformer layers and used for downstream tasks. Assuming that there are N words, the embedding layer input can be expressed as: $W = \{CLS, W_1, W_2, \cdots, W_N\}$. The W, after WordPiece (Sennrich et al., 2016) embedding operation, will get token embeddings. In addition, segment embeddings and position embeddings need to be added to obtain embedding layer output: $L = \{L_{CLS}, L_1, L_2, \cdots, L_N\}$. To keep the same length of the three modalities, add 0 as padding (P) in front of acoustic and visual, respectively. Visual modality can be denoted by $V = \{P, V_1, V_2, \cdots, V_N\}$. And, acoustic modality can be denoted by $A = \{P, A_1, A_2, \cdots, A_N\}$. To bring these modal information together, we prepare a sequence of triplets $[(L_i, V_i, A_i) : \forall_i \in [1, N]]$ by pairing L_i with the corresponding (V_i, A_i). Each triplet will pass through the attachment GFMAM, which is capable of converting each triplet into new multimodal information of the corresponding word embedding. Nonverbal modalities (visual and acoustic) can dynamically adjust the position of words in the semantic space (Wang et al., 2019b). These compact multimodal representations are used to change the position of words in the semantic space. These shifted word representations can be denoted as $L' = \{L_{CLS}, L_1', L_2', \cdots, L_N'\}$. L' is fed into the transformer layers that follow, and the last transformer layer of output $[CLS]$ is used as a label for multimodal sentiment analysis.

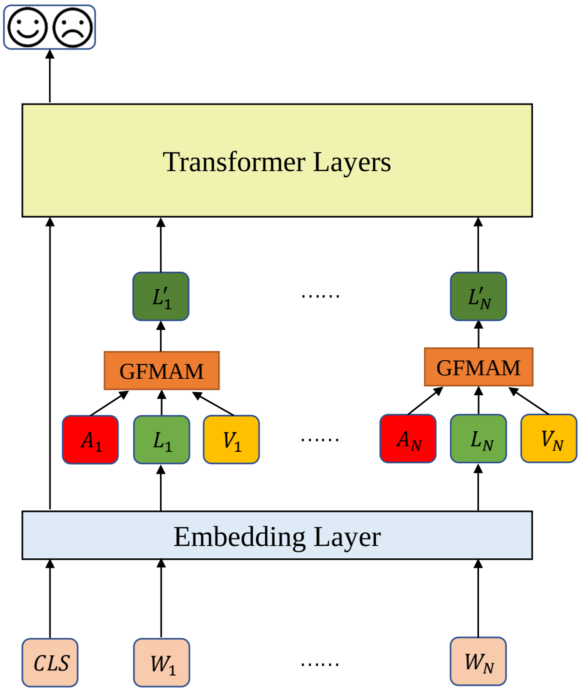

Figure 3: Simplified diagram of the Graph-based Fine-grained Multimodal Attention Mechanism (GFMAM) embedded in the specific location of the BERT.

5 Experiments

This section introduces our experimental settings, including the experimental datasets, evaluations, preprocessing, baselines, results, and analysis.

5.1 Datasets

The proposed algorithm is tested using two public benchmark multimodal sentiment analysis and emotion recognition datasets: CMU-MOSI and CMU-MOSEI. These datasets provide word-aligned multimodal signals (language, visual, and acoustic) for each utterance.

CMU-MOSI: The CMU-MOSI is a commonly used dataset for human multimodal sentiment analysis. It consists of 2,198 short monologue video clips (each clip lasts for the duration of one sentence) expressing the opinion of the speaker inside the video on a topic such as movies. The utterances are manually annotated with a continuous opinion score between $[-3, +3]$, [-3: highly negative, -2 negative, -1 weakly negative, 0 neutral, $+1$ weakly positive, $+2$ positive, $+3$ highly positive].

CMU-MOSEI: The CMU-MOSEI is an improved version of CMU-MOSI. It contains 23,453 annotated video clips (about 10 times more than CMU-MOSI) from 5,000 videos, 1,000 different speakers, and 250 different topics. The number of discourses, samples, speakers, and topics is also larger compared to CMU-MOSI. The range of labels taken for each discourse is consistent with CMU-MOSI.

5.2 Preprocessing

We utilize the standard low-level features that are provided by the respective benchmarks.
Language Modality: Traditionally, language modality features have been GloVe (Pennington et al., 2014) embeddings for each token in the utterance. GloVe features are 300 dimension token embeddings. However, recent works (Rahman et al., 2020; Hazarika et al., 2020) have demonstrated that BERT can provide better performance than GloVe in feature extraction. Therefore, BERT is used to obtain the features of language modality in the proposed method. We utilize the *bert-base-uncased* and *bert-large-uncased* pre-trained models.
Visual Modality: CMU-MOSI and CMU-MOSEI use Facet to extract facial expression features, including facial action units and facial gestures based on a Facial Action Coding System (FACS) (Ekman and Rosenberg, 1997). This process is repeated for each sampled frame within the utterance video sequence. The final visual feature dimensions, d_v, are 47 for CMU-MOSI, 35 for CMU-MOSEI.
Acoustic Modality: COVAREP (Degottex et al., 2014) is used to extract the following relevant features: fundamental frequency, quasi-open quotient, normalized amplitude quotient, glottal source parameters (H1H2, Rd, Rd conf), VUV, MDQ, the first 3 formants, PSP, HMPDM 0-24 and HM-PDD 0-12, spectral tilt/slope of wavelet responses(peak/slope), MCEP 0-24. The final acoustic feature dimension, d_a, is 74 for MOSI/MOSEI.

For each word, we align all three modalities following the convention established in (Chen et al., 2017). Assuming that there are T words in the video, the features for language can be denoted as $T \times d_l$, for visual as $T \times d_v$, and for acoustic as $T \times d_a$.

5.3 Evaluation Criteria

The evaluation metrics of MISA are referred to in the experiments(Hazarika et al., 2020). There are five evaluation metrics, namely Mean Absolute Error (MAE), Pearson Correlation (Corr), Binary Accuracy (Acc-2), F1-Score, and Seven-class Accuracy (Acc-7). MAE and Corr are regression tasks. Acc-2, F1-Score, and Acc-7 are classification tasks. For the calculation of Acc-2, two different evaluation methods are included. The first one is negative/non-negative classification (Zadeh et al., 2018c), where non-negative includes neutral sentiment information. The second one is negative/positive classification (Tsai et al., 2019), excluding neutral sentiments information. The results of all evaluation metrics mentioned above are reported.

5.4 Baselines

The various state-of-the-art models introduced following are used as the baseline for comparison.
TFN: Tensor Fusion Network (TFN) (Zadeh et al., 2017) performs an outer product of the output vectors after encoding the three modes to learn the intra- and inter-modal dynamics in an end-to-end manner and can capture uni-, bi-, and tri-modal interactions.
MFN: Memory Fusion Network (MFN) (Zadeh et al., 2018a) uses three separate LSTMs to model each modality and uses Delta-memory attention and Multi-View Gated Memory to capture both temporal and inter-modal interactions.

Models	CMU-MOSI				
	MAE↓	Corr↑	Acc-2↑	F1-Score↑	Acc-7↑
TFN	0.970	0.633	73.9/-	73.4/-	32.1
MFN	0.965	0.662	77.4/-	77.3/-	34.1
RMFN	0.922	0.681	78.4/-	78.0/-	38.3
MulT	0.871	0.698	-/83.0	-/82.8	40.0
MTAG	0.889	0.686	-/82.1	-/82.3	38.9
MG	0.933	0.684	-/80.6	-/80.5	32.1
TFN(B)[1]	0.901	0.698	-/80.8	-/80.7	34.9
MFN(B)[1]	0.877	0.706	-/81.7	-/81.6	35.4
RMFN(B)[2]	0.878	0.712	79.6/89.7	78.9/79.1	-
MulT(B)[2]	0.861	0.711	81.5/84.1	80.6/83.9	-
ICCN	0.862	0.714	-/83.07	-/83.02	39.01
MISA	0.783	0.761	81.8/83.4	81.7/83.6	42.3
MAG	0.712	0.796	84.2/86.1	84.1/86.0	-
Ours*	**0.689**	**0.809**	**84.3/86.4**	**84.2/86.2**	**48.6**
Human	0.710	0.820	85.7/-	87.5/-	53.9
Ours†	**0.651**	**0.835**	**86.0/88.2**	**86.0/88.2**	**50.6**

Table 1: Performances of multimodal models on CMU-MOSI. Best results are highlighted in bold. NOTE: (B) means the language features are based on BERT; - means the result is not given in the paper; * means the text feature is based on *bert-base-uncased*; † means the text feature is based on *bert-large-uncased*; [1] is from (Sun et al., 2020) and [2] is from (Rahman et al., 2020). Human performance for CMU-MOSI is reported as (Zadeh et al., 2018c). In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive".

Models	CMU-MOSEI				
	MAE↓	Corr↑	Acc-2↑	F1-Score↑	Acc-7↑
TFN(B)[1]	0.901	0.698	-/80.8	-/80.7	34.9
MFN(B)[1]	0.568	0.717	-/84.4	-/84.3	35.4
MG	0.608	0.675	-/81.4	-/81.7	49.7
MulT	0.580	0.703	-/82.5	-/82.3	51.8
ICCN	0.565	0.713	-/84.2	-/84.2	51.6
MISA	0.555	0.756	83.6/85.5	83.8/85.3	52.2
MAG[3]	0.539	0.753	83.7/85.2	83.7/85.0	-
Ours†	**0.517**	**0.786**	**85.2/86.9**	**85.0/86.8**	**54.9**

Table 2: Performances of multimodal models on CMU-MOSEI. Best results are highlighted in bold. NOTE: (B) means the language features are based on BERT; - means the result is not given in the paper; [1] is from (Sun et al., 2020) and [3] is from (Yu et al., 2021).

RMFN: Multimodal Language Analysis with Recurrent Multistage Fusion (RMFN) (Liang et al., 2018) can automatically decompose the multimodal fusion problem into multiple recursive stages. At each stage, a subset of the multimodal signals is highlighted and fused with the previous fusion representation.

MulT: Multimodal Transformer for Unaligned Multimodal Language Sequence (MulT) (Tsai et al., 2019) extends the multimodal transformer architecture by using directional paired cross-attention to transform one modality into another.

ICCN: For Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis, Interaction Canonical Correlation Network (ICCN) (Sun et al., 2020) first extracts features from audio and video modalities, and then fuses them with text embeddings to get two outer products, text-audio, and text-video. Finally, the external products are fed into CCA network, and their output is used to predict.

MG: Analyzing Unaligned Multimodal Sequence via Graph Convolution and Graph Pooling Fusion (MG) (Mai et al., 2020) first uses a graph convolutional network to learn intra-modal dynamics for each modality. Then, a graph pooling fusion network is devised to automatically learn the associations between various nodes from different modalities.

MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA) (Hazarika et al., 2020) combines various losses, including distribution similarity, orthogonal loss, reconstruction loss, and task prediction loss, to learn modality-invariant and modality-specific representations.

MAG: Integrating Multimodal Information in Large Pretrained Transformers (MAGT) (Rahman et al., 2020) is an improved work on RAVEN, which applies Multimodal Adaptive Gate (MAG) on different layers of the BERT backbone.

MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences (Yang et al., 2021) first convert unaligned multimodal sequence data into a graph. Then, an operation called MTAG is designed to capture the various interactions among multimodalities.

For TFN, MFN, RMFN, and MulT, the language features are based on GloVe, while ICCN, MISA, MAG, Self-MM, and our model use language features based on BERT. For the sake of fairness, we also provide the results of these models using BERT to obtain language features.

5.5 Comparison with Baselines

Table 1 shows the results of our model in comparison with other models and humans on the CMU-MOSI dataset. It can be observed that the model proposed in this paper works better, and all the evaluation metrics are better than other models. As compared to other work that relies on graph

Data View	CMU-MOSI				
	MAE↓	Corr↑	Acc-2↑	F1-Score↑	Acc-7↑
No Compensate	0.652	0.820	85.7/88.2	85.7/88.2	50.2
Compensate	0.651	0.835	86.0/88.2	86.0/88.2	50.6

Table 3: Results for experiments on CMU-MOSI. We compare the best results obtained with and without the compensation operation on the CMU-MOSI dataset.

neural networks, like MTAG and GM, our model exhibits excellent performance. The reason is that our model takes into account the use of information from different subspaces to construct tri-modal interactions. And a text-dominant multimodal fusion scheme is designed. Also, Zadeh et al. (2018c) reported human performance results on the CMU-MOSI dataset. We can observe that performance results outperform human performance in binary classification (Acc-2) and regression tasks (MAE, Corr). To the best of our knowledge, the accuracy of binary classification exceeds that of humans for the first time.

Table 2 shows the performance results of our model on the CMU-MOSEI dataset, where all evaluation metrics outperform the other models. The performance exhibited by our model validates the usefulness of constructing multimodal interactions at the subspace level. Based on the evaluation results of two publicly available datasets, our proposed model is successful for multimodal sentiment analysis.

5.6 Ablation Studies

To verify the effectiveness of the proposed model, the following ablation experiments are designed. There are two questions.

Question 1: Is it useful to compensate for the continuity within the modalities?

Question 2: Is the splitting processing effective for multimodal sentiment analysis?

For **Question 1**, we compare our proposed model on the CMU-MOSI dataset in different cases. These cases include the model with and without continuity complement. As shown in Table 3, the result of the model with complement is a little better than the model without complement. The experiment results demonstrate that although the improvement of the complement is not obvious, the information of the modality continuity is disrupted by the splitting operation. Therefore, the complement of internal continuity is required when the splitting operation is performed on modal information.

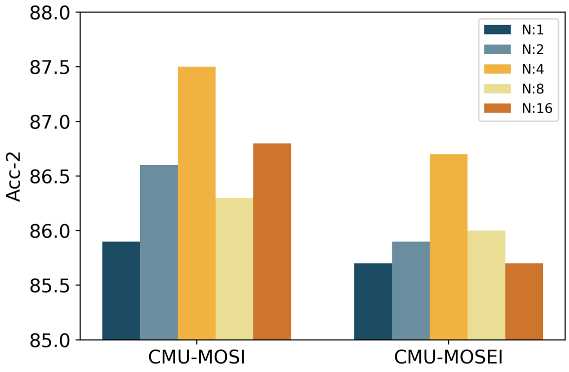

Figure 4: Results for experiments on CMU-MOSI and CMU-MOSEI. The performance exhibited by our model under different fine-grained multimodal information. N denotes the number of splits for each modality.

For **Question 2**, during the experiments, we split multimodal information into different granularities, keeping the rest of the hyper-parameters constant. Figure 4 shows the results of the binary classification obtained after splitting each modality for the CMU-MOSI and CMU-MOSEI datasets, where N denotes the number of nodes. And when N=1, it means that the modal information will not be split, that is, the original sampling rate will be maintained. The cases of N>1 are significantly better than those for N=1. The results validate that integrating multimodal information at the subspace level can improve the performance of sentiment analysis.

6 Conclusion

In this paper, we propose a novel GFMAM attachment, which can effectively fuse fine-grained multimodal information at the subspace level for sentiment analysis. Without changing the architecture of the original BERT, the fine-grained multimodal information is effectively fused with the graph structure. Furthermore, we demonstrate that multimodal information is necessary for fine-grained interactions by conducting ablation studies in our models. The experimental results demonstrate the effectiveness of the proposed method when doing sentiment analysis tasks and show the best performance on public datasets.

In the future, the fine-grained multimodal interactions across multiple moments will be considered to further improve the performance of the sentiment analysis.

References

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 163–171.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492.

Sijie Mai, Songlong Xing, Jiaxuan He, Ying Zeng, and Haifeng Hu. 2020. Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. *arXiv preprint arXiv:2011.13572*.

Sijie Mai, Songlong Xing, and Haifeng Hu. 2021. Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1424–1437.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

Zhongkai Sun, Prathusha K Sarma, Yingyu Liang, and William Sethares. 2021. A new view of multi-modal language analysis: Audio and video features as text "styles". In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1956–1965.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019a. Heterogeneous graph attention network. In *The World Wide Web Conference*, pages 2022–2032.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019b. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1437–1445.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Jianfeng Wu, Sijie Mai, and Haifeng Hu. 2021. Graph capsule aggregation for unaligned multimodal sequences. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 521–529.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer. 2018b. Proceedings of grand challenge and workshop on human multimodal language (challenge-hml). In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018c. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018d. 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)*, pages 2236–2246.