electronics

MDPI

Article

Dynamic Weighted Multitask Learning and Contrastive Learning for Multimodal Sentiment Analysis

Xingqi Wang [1,2], Mengrui Zhang [1], Bin Chen [1,*], Dan Wei [1] and Yanli Shao [1]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[2] Key Laboratory of Discrete Industrial Internet of Things of Zhejiang Province, Hangzhou 310018, China
* Correspondence: chenbin@hdu.edu.cn; Tel.: +86-181-5712-6472

Abstract: Multimodal sentiment analysis (MSA) has attracted more and more attention in recent years. This paper focuses on the representation learning of multimodal data to reach higher prediction results. We propose a model to assist in learning modality representations with multitask learning and contrastive learning. In addition, our approach obtains dynamic weights by considering the homoscedastic uncertainty of each task in multitask learning. Specially, we design two groups of subtasks, which predict the sentiment polarity of unimodal and bimodal representations, to assist in learning representation through a hard parameter-sharing mechanism in the upstream neural network. A loss weight is learned according to the homoscedastic uncertainty of each task. Moreover, a training strategy based on contrastive learning is designed to balance the inconsistency between training and inference caused by the randomness of the dropout layer. This method minimizes the MSE between two submodels. Experimental results on the MOSI and MOSEI datasets show our method achieves better performance than the current state-of-the-art methods by comprehensively considering the intramodality and intermodality interaction information.

Keywords: multimodal sentiment analysis; multitask learning; contrastive learning

Citation: Wang, X.; Zhang, M.; Chen, B.; Wei, D.; Shao, Y. Dynamic Weighted Multitask Learning and Contrastive Learning for Multimodal Sentiment Analysis. *Electronics* **2023**, *12*, 2986. https://doi.org/10.3390/electronics12132986

Academic Editor: Ping-Feng Pai

Received: 21 June 2023
Revised: 3 July 2023
Accepted: 4 July 2023
Published: 7 July 2023

1. Introduction

With the development of social media, netizens have begun to make comments and expressed views by utilizing diverse modalities other than text, such as audio and video. Analyzing the emotional information embedded in these multimodal messages has become crucial for market analysis, preference analysis, and other related fields. Consequently, multimodal sentiment analysis (MSA) has gained growing attention in recent years [1–3].

Compared to the conventional text-specific sentiment analysis tasks, MSA incorporates two or more modalities as inputs. MSA achieves superior accuracy in sentiment prediction by integrating various pieces of modal information, including natural language, facial expressions, and voice intonation. Figure 1 illustrates that leveraging multiple modalities provides significant advantages over relying solely on data from a single modality for sentiment analysis.

The five main challenges of multimodal tasks are representation, translation, alignment, fusion, and colearning [4]. Multimodal fusion is one of the most important topics in multimodal learning, which can be categorized into three types: early, late, and hybrid fusion [5]. A significant research focus in multimodal fusion is how to extract effective complementary information from multiple modalities and integrate it into a fused representation. Zadeh et al. [6] introduced the tensor fusion network (TFN), which captured interactions within and between modalities by computing tensor cross-products of modalities. Liu et al. [7] proposed a low-rank multimodal fusion approach, leveraging low-rank tensors to reduce the computational complexity of tensor methods while performing multimodal fusion. Hu et al. [8] proposed a multimodal sentiment knowledge-sharing

framework (UniMSE) which fused modalities at the syntactic and semantic levels and incorporated contrastive learning to capture the consistency and differences between sentiments and emotions.

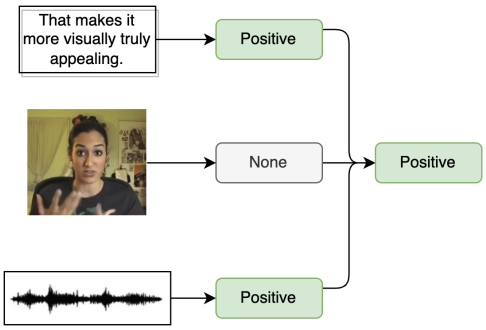


Figure 1. Multimodal sentiment analysis based on text, audio, and video.

Based on the review of previous works, it is observed that most studies employ simple traditional neural networks to extract modal vectors, which are then directly used as inputs for the representation fusion module. These works primarily emphasize representation fusion, while potentially ignoring the significance of learning modal representations. The fusion of representations can sometimes suppress the predictive power of individual modalities, despite the fact that each piece of unimodal data contains rich sentiment information. Therefore, leveraging the hidden information in these heterogeneous data sources can contribute to training an effective model and enhancing the prediction accuracy. Tsai et al. [9] proposed a joint generation–discriminant objective optimization method, which decomposed the representation into multimodal discriminants and mode-specific generation factors. The former was employed for sentiment classification, while the latter facilitated the learning of mode-specific generation features. Sun et al. [10] introduced the interaction canonical correlation network (ICCN), which learned multimodal embeddings by capturing hidden correlations among different modalities. In the context of multimodal data where random modality data may be missing, Sun et al. [11] proposed a framework called efficient multimodal transformer with dual-level feature restoration (EMT-DLFR) to enhance the robustness of models in such scenarios. EMT utilized utterance-level representations of each modality as global multimodal context and interacted with local unimodal features, thereby encouraging the model to learn semantic information from incomplete data.

In this paper, we introduce a model named MCM (**m**ultitask learning and **c**ontrastive learning for **m**ultimodal sentiment analysis) to assist in learning modal representations. Our proposed model comprises two key components: a multitask learning module and a contrastive learning module. Given the diverse nature of multimodal data, it presents an excellent opportunity for leveraging multitask learning. To exploit this potential, we design subtasks specific to different modal representations, aiming to effectively extract the underlying emotional information. While multitask learning primarily serves as a method for unimodal representation learning, we extend our investigation to the learning of fusion representations. We achieve this by incorporating contrastive learning to constrain the fusion prediction, enabling the assisted learning of fusion representations. This approach adds additional complexity to our model, allowing for a more comprehensive learning of the multimodal representations.

In previous studies, multitask learning was applied to multimodal sentiment analysis tasks, and a common characteristic of these studies was that they only utilized unimodal data as subtasks for auxiliary learning [12]. However, we believe that such a design tends to excessively focus on modeling within each modality while overlooking the modeling of interactions between modalities. The key distinction between multimodal and unimodal tasks lies in the interactions among different modalities. Therefore, in addition to the subtasks targeting unimodal data, we propose incorporating subtasks specifically designed

for bimodal representations generated by a gating mechanism. This design allows our model to simultaneously consider both intramodal and intermodal interactions, fully harnessing the advantages of multimodal tasks.

Contrastive learning was initially used in unsupervised learning tasks to learn sentence embeddings from unsupervised data, and later it was gradually extended to supervised data [13]. Previous research has demonstrated that dropout can lead to inconsistencies between the training and inference stages [14,15], which can have a detrimental effect on the final multimodal sentiment polarity prediction results. Therefore, the randomness of the dropout mechanism can be effectively utilized to maintain an output consistency by constraining multiple prediction results, thereby enhancing the overall performance of the model.

For the multitask module in MCM, the main task focuses on predicting the sentiment polarity of fusion representations, while the subtasks involve predicting the sentiment polarity of unimodal representations extracted by traditional networks and bimodal representations generated through a gating unit. By jointly training these tasks, we aim to capture the sentiment information hidden in the vectors. To enhance the performance of multitask learning, we propose a strategy that dynamically acquires task weights based on the existence of homoscedastic uncertainty [16] in the data. This approach replaces the conventional manual weight setting method, improving the predictive results and reducing computation time. For the contrastive learning module in MCM, we mitigate the issue of training and inference inconsistency caused by the randomness of the dropout mechanism by constraining the consistency of the output results from the two submodels.

The contributions of our work can be summarized as follows:

1. We propose a dynamic weighted multitask learning method to facilitate the learning of hidden emotional information within modal representations. By assigning dynamic weights based on homoscedastic uncertainty, our approach enhances the effectiveness of multitask learning.
2. Our method incorporates a contrastive learning module, which ensures the consistency between training and inference by constraining the training of the model. This module optimizes the training process and improves the overall performance of the model.
3. Experimental results on two widely used datasets, MOSI and MOSEI, demonstrate the superiority of our method compared to current approaches in the field of multimodal sentiment analysis. Our approach achieves comprehensive representation learning under the consideration of both intramodal and intermodal interactions, resulting in improved performance.

2. Related Work

2.1. Multimodal Sentiment Analysis

Multimodal sentiment analysis is a multidisciplinary research field that encompasses natural language processing, computer vision, speech processing, and more. Zadeh et al. [17] introduced the memory fusion network (MFN), a multiview gated memory network that captures both in-view and cross-view interactions. Tsai et al. [18] proposed the cross-modal transformers model (MulT), which strengthened the target modality through cross-modal attention learning. In the late fusion stage, MulT first learned intramodal representations and then performs intermodal fusion. Hazarika et al. [19] proposed MISA, a method capable of learning modality-invariant and modality-specific representations. Rahman et al. [20] proposed a method that achieved data alignment by employing multimodal adaptive gates in different layers of BERT [21] and XLNet [22]. Han et al. [23] presented the MultiModal InfoMax (MMIM) framework, which maximized mutual information hierarchically during multimodal fusion. MMIM enhanced the mutual information between unimodal input pairs, as well as between multimodal fusion results and unimodal input.

Xue et al. [24] proposed a multi-level attention map network (MAMN) to address the denoising problem within and between multimodal inputs. The MAMN filtered out noise in the fusion process and captured both consistent and heterogeneous correlations across multiple granularities. MAMN consisted of three modules: the multigranularity feature extraction module, the multilevel attention map generation module, and the attention map fusion module. Considering that traditional methods struggle to capture the global contextual information of long time series data when extracting temporal features from a single modality and often overlook the correlations between modalities during multimodal fusion, Cheng et al. [25] proposed the attentional temporal convolutional network (ATCN) for extracting temporal features from individual modalities. They also introduced the multilayer feature fusion (MFF) model, which utilized different methods to fuse features at various levels based on their correlation coefficients, thereby enhancing the effectiveness of the multimodal fusion.

Wang et al. [26] proposed the recurrent attended variation embedding network (RAVEN), which focused on learning word representations by modeling the fine-grained structure of nonverbal modalities. It is worth mentioning that RAVEN employs an attention mechanism to calculate the shift vector of the text representation, enabling the nonlinear combination of visual and acoustic embedding. Yu et al. [12] proposed a model called Self-MM, which jointly trained unimodal and multimodal representations to capture consistency and differences. The unimodal labels in Self-MM were obtained using a label generation module based on self-supervised learning. Different from Self-MM, our approach includes a set of bimodal subtasks in addition to the unimodal subtasks. Furthermore, we directly use multimodal sentimental intensity labels as targets for subtask learning, aiming to learn representations that comprehensively consider both intramodal and intermodal interactions by aligning the training results of the subtasks with the multimodal learning targets.

2.2. Multitask Learning

Multitask learning is one of the transfer learning methods, which aims to leverage valuable information contained in multiple related tasks to enhance the generalization performance of all tasks [27]. There are two main mechanisms for parameter sharing in multitask learning: hard parameter sharing and soft parameter sharing. The hard parameter sharing mechanism [28], illustrated in Figure 2a, is the most commonly used parameter sharing strategy. In this approach, multiple tasks share the parameters of the upstream network while maintaining their own independent task-specific output layers to achieve specific tasks. This parameter sharing method enables downstream tasks to learn more comprehensive and informative representations from the shared upstream layer. The soft parameter sharing mechanism [29], as shown in Figure 2b, does not directly share network parameters among tasks. Instead, each task has its own independent network with a dedicated set of parameters. The parameter space sharing is achieved by imposing distance regularization constraints on each model parameter.

In this work, we employed the hard parameter sharing method. Specifically, two sets of subtasks shared partial upstream network parameters with the main task. This parameter sharing method allowed the subtasks to benefit from the shared network while focusing on their specific objectives. Furthermore, we introduced a dynamic weighted method based on homoscedasticity uncertainty. This approach allowed us to assign dynamic weights to the subtasks during training, based on their respective uncertainties. By incorporating homoscedasticity uncertainty, we can adaptively adjust the importance of each subtask and optimize the overall learning process.

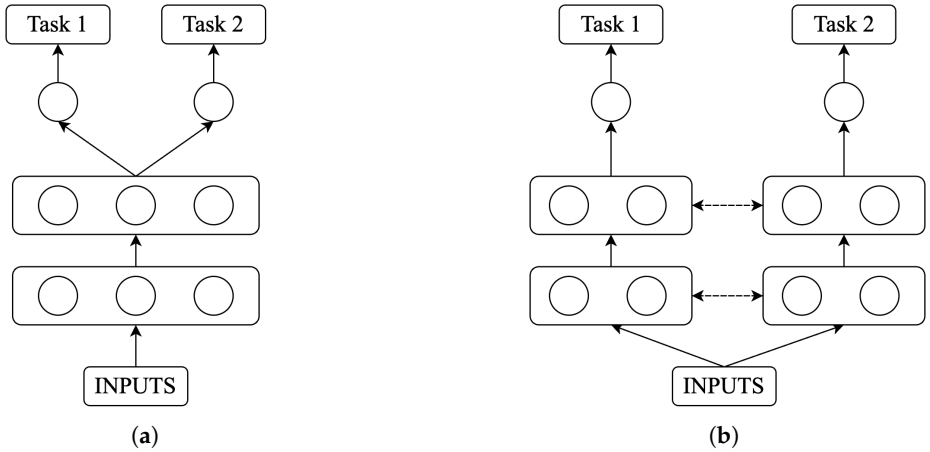


Figure 2. Sharing mechanisms of multitask learning. (**a**) Hard parameter sharing; (**b**) Soft parameter sharing.

2.3. Contrastive Learning

Contrastive learning is a type of self-supervised learning that aims to learn effective representation by bringing semantically similar samples closer together while pushing semantically dissimilar samples apart [30]. An example of contrastive learning is the SimCSE method proposed by Gao et al. [31], which focuses on learning sentence vectors by using contrastive learning method. SimCSE leverages the randomness of dropout [32] for data augmentation, mitigating the concern that manual data augmentation might alter the semantics of the data. It is important to note that SimCSE is specifically designed for unsupervised tasks.

With the advancement of research on contrastive learning, it has gradually been applied to supervised tasks. Wu et al. [13] proposed a method called R-Drop, which built upon the idea of "Dropout Twice" similar to SimCSE. R-Drop suggested that two distinct submodels could be obtained by applying dropout twice. The final loss was computed by combining the predicted losses of these two submodels with the Kullback–Leibler (KL) divergence of their different outputs after applying dropout twice. This approach leveraged the randomness of the dropout mechanism to create diverse submodels and encouraged consistency between the predictions of these submodels, leading to enhanced representation learning and improved performance in supervised tasks.

The randomness of the dropout mechanism can introduce inconsistencies between the training and prediction phases. During training, dropout randomly deactivates certain parameters in the neural network by setting them to zero. Consequently, determining the optimal model within the ensemble becomes challenging. In theory, averaging the multiple predictions of the same input is a reasonable approach to obtain the final prediction for the models with a dropout layer. However, in practical prediction scenarios, dropout is typically disabled, and no parameters are set to zero, leading to inconsistencies between training and inference. To address this issue, we incorporated the contrastive learning strategy, which introduced a regularization term that encouraged consistency among the model's outputs under different dropout configurations. Specifically, we applied the concept of "Dropout Twice" to the task-specific output layer of the main prediction task following the multimodal representation fusion layer. The mean squared error (MSE) between the outputs of the two submodels was set as the regularization term. This approach effectively addresses the challenge of inconsistency between training and inference. This regularization enhances the model's robustness and improves its overall performance.

3. Methodology

In this section, we provide an overview of the MCM model. We first introduce the general structure of MCM, followed by a detailed description of the dynamic weighted multitask learning module and the contrastive learning module incorporated within MCM.

3.1. Overall Architecture

The architecture of the proposed MCM model in this paper is illustrated in Figure 3. Our model comprises three sets of tasks, with multimodal sentiment analysis as the main task and unimodal and bimodal sentiment analysis as subtasks.

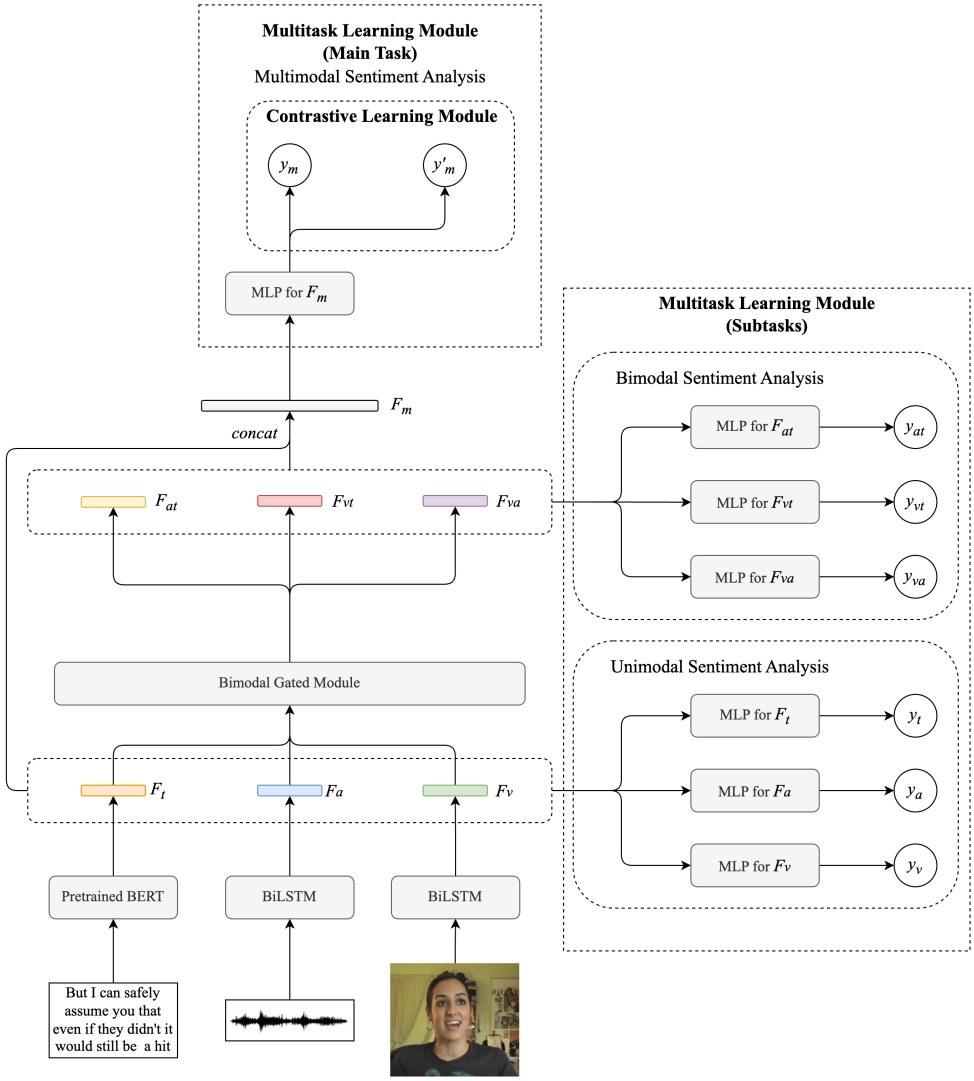


Figure 3. The architecture of MCM.

Unimodal Sentiment Analysis. The original representations of text, audio, and video are denoted as $f_t \in \mathbb{R}^{d_t}$, $f_a \in \mathbb{R}^{d_a}$, and $f_v \in \mathbb{R}^{d_v}$, respectively. d denotes the dimension of the representation, and θ denotes the trainable parameters in the corresponding network. For the text modality, we extracted features using a pretrained BERT model. The vector F_t, which corresponds to the embedding of the first word in the output of the last layer of BERT, was taken as the representation of the whole sentence.

$$F_t = BERT(f_t; \theta_t) \in \mathbb{R}^{d_t} \quad (1)$$

For the audio and video modalities, we extracted features by using a bidirectional LSTM network [33] to capture the sequential characteristic of the audio and video data.

$$F_a = LSTM(f_a; \theta_a) \in \mathbb{R}^{d_a} \tag{2}$$

$$F_v = LSTM(f_v; \theta_v) \in \mathbb{R}^{d_v} \tag{3}$$

The final prediction part consisted of two linear transformations followed by a ReLU activation function. W denotes the weights in the linear layer, while b denotes the bias parameters.

$$F_s^{'} = ReLU(W_{u1}^s F_s + b_{u1}^s) \tag{4}$$

$$y_s = W_{u2}^s F_s^{'} + b_{u2}^s \tag{5}$$

where $s \in \{t, a, v\}$, $W_{u1}^s \in \mathbb{R}^{d_s \times d_{s'}}$, $W_{u2}^s \in \mathbb{R}^{d_{s'} \times 1}$, and y_s is the prediction result of the unimodal task. We took these three unimodal prediction tasks as the first group of subtasks. The upstream network and parameters for learning unimodal representations were shared with the main task.

Bimodal Sentiment Analysis. The purpose of the gating mechanism is to generate an intermediate representation by combining data from different modalities [34]. To achieve this, we designed a bimodal gated module that learned bimodal representations with dimension h by integrating the information from two unimodal representations. The module structure is shown in Figure 4.

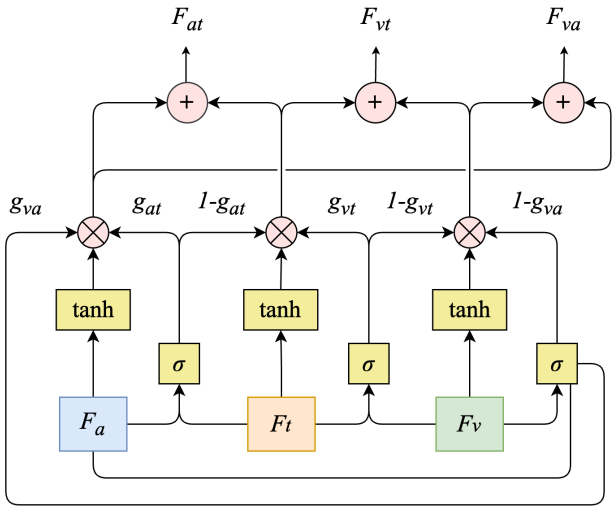


Figure 4. Bimodal gated module.

Specifically, first we combined the three unimodal representations pairwise as F_α, F_β, where $(\alpha, \beta) \in \{(a, t), (v, t), (v, a)\}$. These combined representations were then taken as the input for linear layers with a tanh activation function.

$$h_\alpha = tanh(W_{b1}^{\alpha\beta} F_\alpha + b_{b1}^{\alpha\beta}) \tag{6}$$

$$h_\beta = tanh(W_{b2}^{\alpha\beta} F_\beta + b_{b2}^{\alpha\beta}) \tag{7}$$

where $W_{b1}^{\alpha\beta} \in \mathbb{R}^{d_\alpha \times h}$, $W_{b2}^{\alpha\beta} \in \mathbb{R}^{d_\beta \times h}$. We concatenated the two representations F_α and F_β. Then, $g_{\alpha\beta}$, a weight controlling the contribution of two unimodal inputs, was calculated by a linear transformation and a ReLU activation function.

$$g_{\alpha\beta} = ReLU(W_{b3}^{\alpha\beta}[F_\alpha; F_\beta] + b_{b3}^{\alpha\beta}) \tag{8}$$

where $W_{b3}^{\alpha\beta} \in \mathbb{R}^{(d_\alpha+d_\beta)\times h}$. Finally, the bimodal representation $h_{\alpha\beta}$ was calculated by a weighted sum.

$$F_{\alpha\beta} = g_{\alpha\beta}h_\alpha + (1 - g_{\alpha\beta}h_\beta) \tag{9}$$

The final prediction part consists of two linear transformations and a ReLU activation function.

$$F'_{\alpha\beta} = ReLU(W_{b4}^{\alpha\beta}F_{\alpha\beta} + b_{b4}^{\alpha\beta}) \tag{10}$$

$$y_{\alpha\beta} = W_{b5}^{\alpha\beta}F'_{\alpha\beta} + b_{b5}^{\alpha\beta} \tag{11}$$

where $W_{b4}^{\alpha\beta} \in \mathbb{R}^{h\times h'}$, $W_{b5}^{\alpha\beta} \in \mathbb{R}^{h'\times 1}$, and $y_{\alpha\beta}$ denotes the prediction result of the bimodal task. These three bimodal prediction tasks were taken as the second group of subtasks. Similar to unimodal sentiment analysis, the upstream layers used to learn bimodal representations were shared with the main task.

Multimodal Sentiment Analysis. As the main task of MCM, this task combines the unimodal and bimodal representations and utilizes the fusion representations as input for the multimodal sentiment prediction network. In addition to the representation learning layers shared with unimodal and bimodal sentiment prediction, this task includes a task-specific sentiment polarity prediction layer to get the final results.

In the first stage, the multimodal fusion representation F_m was constructed by concatenating three unimodal and three bimodal representations.

$$F_m = [F_t; F_a; F_v; F_at; F_vt; F_va] \tag{12}$$

In the second stage, the multimodal prediction result was derived through a linear regression.

$$F'_m = ReLU(W_1^m F_m + b_1^m) \tag{13}$$

$$y_m = W_2^m F'_m + b_2^m \tag{14}$$

where $W_1^m \in \mathbb{R}^{(d_t+d_a+d_v+3\times h)\times d_m}$, $W_2^m \in \mathbb{R}^{d_m\times 1}$, and y_m denotes the prediction result of the multimodal task. It represents the final sentiment analysis prediction result of MCM.

3.2. Multitask Learning Module

In this paper, the independent sentiment polarity prediction of three unimodal representations F_t, F_a, F_v and three bimodal representations h_{at}, h_{vt}, h_{va} were considered as two groups of subtasks. The prediction of the multimodal fusion representation F_m was taken as the main task. By the joint training of multiple tasks, the shared layers between the main task and subtasks were trained simultaneously. The visual sharing graph between different tasks in MCM is shown in Figure 5. From the graph, we can intuitively observe that the shared layers between the unimodal subtasks and the main task were designed to capture the emotional information present in the unimodal data and generate unimodal representations. This design enabled the model to capture underlying sentimental information within the unimodal data, thereby enhancing the effectiveness of the generated unimodal representations.

However, relying solely on unimodal subtasks may cause certain limitations. In the case of multimodal tasks, it is crucial to consider both intramodal and intermodal interactions. Training auxiliary tasks for individual modalities primarily focuses on capturing intramodal interactions, neglecting intermodal interactions. Consequently, the unimodal representations obtained through this approach may exhibit stronger independent predictive capabilities, but they may not be optimal for subsequent modal fusion stages. To address this, we introduced bimodal prediction subtasks. The bimodal representations derived from the gating unit could effectively capture intermodal information, thereby compensating for the limited learning of intermodal interactions in the unimodal subtasks. Thus, the bimodal subtasks facilitated the learning of intermodal interaction information, serving as valuable support for multimodal tasks. Additionally, they acted as constraints

on the unimodal subtasks, preventing the learned unimodal representations from deviating too far from the requirements of the multimodal task.

The loss of each task was calculated by the mean squared error (MSE). The simple loss function of multitask learning was calculated as follows:

$$L_{MT} = \sum_{i \in k} (\frac{1}{N} \sum_{j=1}^{N} (\sigma_i \| y_{ij} - \hat{y} \|^2)) \tag{15}$$

where $k \in \{m, t, a, v, ta, tv, va\}$. σ represents the weight coefficient for each task and is a hyperparameter. y_i represents the predicted sentiment intensity score for each task, and $\hat{y}$ represents the sentiment intensity label. The weight coefficient σ described above is usually set manually, which is inaccurate and time-consuming. Thus, we proposed a method to weigh the loss function by considering the homoscedastic uncertainty of each task.

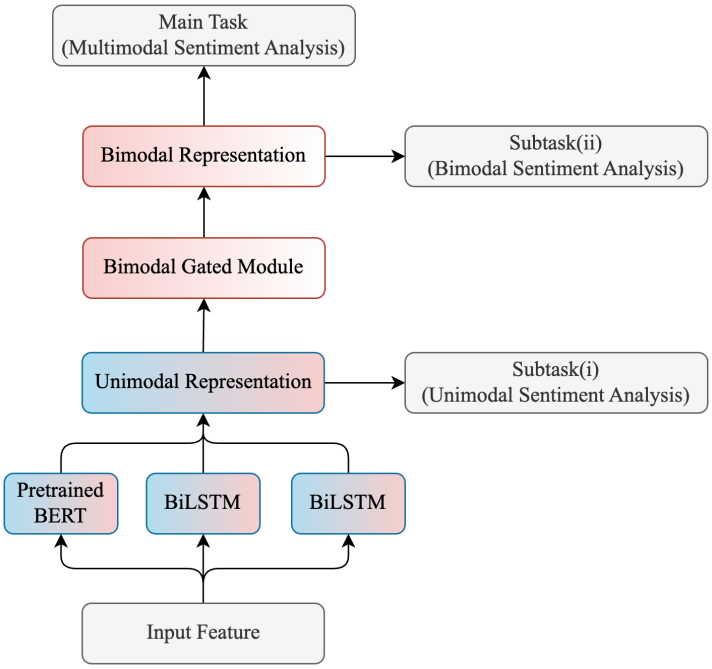


Figure 5. The visual sharing graph between different tasks in MCM. The blocks with blue color indicate that this part of the network structure is shared between the unimodal subtasks and the main task, while the red blocks indicate that this part of the network and parameters are shared between the bimodal subtasks and the main task.

There are two types of uncertainties commonly observed in deep learning: epistemic uncertainty and aleatoric uncertainty [35]. Epistemic uncertainty arises from a lack of training samples, while aleatoric uncertainty arises from unexplained information in the training data and can be further categorized into data-dependent heteroscedastic uncertainty and task-dependent homoscedastic uncertainty [36]. The former depends on the input data, while the latter depends on different tasks. Both types of uncertainty can be captured using Bayesian deep learning methods [37]. In this paper, we addressed the heteroscedastic uncertainty by incorporating a dynamic weighted multitask learning loss function based on a Bayesian neural network. This approach allowed us to effectively consider the uncertainty associated with different tasks and optimize the model accordingly.

Define $f^w(x)$ as the final output of the neural network when the input is x with weight w. For regression tasks, we defined a Gaussian likelihood with a noise scalar σ:

$$p(y|f^w(x)) = \mathcal{N}(f^w(x), \sigma^2) \tag{16}$$

For a multitask learning model with multiple outputs, we defined $f^w(x)$ as the sufficient statistics. For k outputs $y_1, \ldots, y_k$, the multitask likelihood was defined as follows:

$$p(y_1, \ldots, y_k | f^w(x)) = p(y_1 | f^w(x)) \ldots p(y_k | f^w(x)) \tag{17}$$

The negative log likelihood was calculated as follows:

$$-\log p(y_1, \ldots, y_k | f^w(x)) \propto \sum_{i=1}^{k} (\frac{1}{2\sigma_i^2} \|y_i - f^w(x)\|^2 + \log \sigma_i) \tag{18}$$

The optimization objective, which served as the loss function for multitask learning, was defined based on the maximum likelihood estimate.

$$L_{MT} = -\log p(y_1, \ldots, y_k | f^w(x)) \propto \sum_{i=1}^{k} (\frac{1}{2\sigma_i^2} \|y_i - f^w(x)\|^2 + \log \sigma_i) \tag{19}$$

Therefore, the multitask learning loss function in this paper was modified as follows:

$$L_{MT} = \sum_{i \in k} (\frac{1}{N} \sum_{j=1}^{N} (\frac{1}{2\sigma_i^2} \|y_{ij} - \hat{y}\|^2 + \log \sigma_i)) \tag{20}$$

σ is a dynamic adaptive parameter. y_{ij} is the output of modality k, where $k \in \{m, t, a, v, at, vt, va\}$, for input x. $\hat{y}$ is a truth label. The ground truth labels used by all the tasks in this paper were multimodal sentiment intensity labels provided by the dataset.

3.3. Contrastive Learning Module

For the task of predicting the sentiment polarity for fusion representations, the multimodal fusion vector F_m was passed into the network twice to get two sets of prediction results. Subsequently, we compute the mean squared error (MSE) between these two sets of results, which serves as the loss function for the contrastive learning module.

$$L_{CL} = \frac{1}{N} \sum_{j=1}^{N} \|y_{mj} - y'_{mj}\|^2 \tag{21}$$

where N is the size of the dataset, and y_{mj} and y'_{mj} represent the two predicted sentiment intensity scores obtained after applying dropout twice, respectively. The MSE between the output of two submodels was utilized as the optimal objective, aiming to minimize the discrepancy between the two predictions and encourage the model to generate consistent and reliable results. The contrastive learning module played a crucial role in enhancing the alignment and coherence of the predictions, thereby improving the overall performance of the multimodal sentiment analysis task.

3.4. Optimization Objectives

By incorporating both multitask learning and contrastive learning into the training objective, the final loss function was defined as follows:

$$L = L_{MT} + \alpha L_{CL} = \sum_{i \in k} (\frac{1}{N} \sum_{j=1}^{N} (\frac{1}{2\sigma_i^2} \|y_{ij} - \hat{y}\|^2 + \log \sigma_i)) + \alpha \frac{1}{N} \sum_{j=1}^{N} \|y_{mj} - y'_{mj}\|^2 \tag{22}$$

where $k \in \{m, t, a, v, at, vt, va\}$. α represents a hyperparameter.

4. Experiments

4.1. Dataset

In this paper, our model was evaluated on two open multimodal sentiment analysis datasets, MOSI [38] and MOSEI [39].

MOSI. CMU-MOSI consists of 2199 short videos edited from 93 monologue movie commentary videos available on YouTube. The dataset is divided into 1284 training samples, 229 validation samples, and 686 test samples. Each sample is manually annotated by human annotators with a sentiment score ranging from -3 to 3. A higher score indicates a stronger positive emotion, while a lower score indicates a stronger negative emotion.

MOSEI. CMU-MOSEI consists of 23,453 annotated sentences collected from videos featuring over 1000 online speakers discussing 250 different topics on YouTube. The dataset is divided into 16,326 training samples, 1871 validation samples, and 4659 test samples. Similar to MOSI, each sample in CMU-MOSEI is annotated with a sentiment score ranging from -3 to 3, representing the intensity of the sentiment expressed in the sentence.

4.2. Feature Extraction

Text. The raw text data were obtained by manually transcribing the utterances from the video sources. To extract sentence-level text features, we utilized the BERT pretrained model. This choice was motivated by the fact that BERT had undergone extensive pretraining on a large corpus and had demonstrated excellent feature capturing and language representation capabilities. BERT's pretraining involves two tasks, namely masked language model (MLM) and next sentence prediction (NSP). The resulting pretrained model is able to generate text features with a dimension of 768, which is consistent with both datasets under consideration.

Audio. More than 32 audio features, including NAQ (normalized amplitude quotient), MFCCs (Mel-frequency cepstral coefficients), peak slope, energy slope, were extracted using the COVAREP toolkit [40]. The dimension of the audio features was 5 for the MOSI dataset and 74 for the MOSEI dataset. These features provided valuable information about the acoustic characteristics of the utterances, enabling the model to capture audio-related cues for sentiment analysis.

Video. The video features were extracted using Facet, a tool that captures facial expression features for each frame. These features include 16 facial action units, 68 facial landmarks, head pose and orientation, 6 basic emotions, and eye gaze [41,42], based on a facial action coding system. The dimension of the video features was 20 for the MOSI dataset and 35 for the MOSEI dataset. These features, which captured facial expressions associated with emotions, played a crucial role in enabling the model to perform sentiment analysis effectively.

4.3. Baselines

We compared the performance of MCM with the following baseline methods.

TFN (tensor fusion network) obtains unimodal, bimodal, and trimodal interaction information by calculating the outer product of the multimodal tensor.

LMF (low-rank multimodal fusion) is an improvement of the TFN that uses a low-rank tensor for the multimodal fusion, reducing the computational complexity of tensor-based methods.

MFN (memory fusion network) is a multiview sequential gated memory network that models view-specific and cross-view interactions.

RAVEN (recurrent attended variation embedding network) is a method that assists in learning word embeddings by modeling the fine-grained structure of nonverbal modalities.

MFM (multimodal factorization model) learns modality-specific generative features and discriminative features for classification by decomposing representations into generative and discriminative factors.

MulT (multimodal transformer) proposes a multimodal transformer structure that captures the interactions between different multimodal sequences by using bidirectional cross-modal attention.

Self-MM implements joint learning for unimodal and multimodal to learn the consistency and differences between different modal representations.

4.4. Basic Settings

Experimental Design. We use Adam as the optimizer with learning rates of $\{1 \times 10^{-3}, 1 \times 10^{-4}, 5 \times 10^{-5}\}$. The hyperparameter α was set to one. The dimension h of the three bimodal representations was unified as 512. To ensure the robustness of our results, we ran our model five times with different random seeds within $\{1111, 1112, 1113, 1114, 1115\}$ for each task. The final result was obtained by averaging the outcomes of these five runs.

Evaluation Metrics. In line with previous works, we employed four evaluation metrics to assess the effectiveness of our method. Specifically, we utilized the Acc-2 (binary classification accuracy) and F1 score to evaluate the classification performance. The Acc-2 metric measures the percentage of correctly classified samples out of the total number of samples, which is evaluated in two ways: negative/non-negative [43], and negative/positive [18]. The former considers zero as a negative sentiment intensity, while the latter excludes zero from the classification, focusing solely on nonzero sentiment intensities. The F1 score is calculated in the same two ways, and its calculation formula is as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{23}$$

Precision refers to the percentage of samples that are correctly predicted as "positive" out of all samples predicted as "positive". Recall refers to the percentage of samples predicted as "positive" out of all samples that are actually "positive". F1 score can effectively evaluate the datasets with imbalanced samples. The F1 score is a metric that combines precision and recall and is particularly useful for evaluating datasets with imbalanced samples. It provides a balanced measure of the model's performance by considering both precision and recall.

In addition, we used the MAE (mean absolute error) and Corr (Pearson correlation coefficient) to assess the regression performance. The MAE measures the average absolute difference between the predicted values and the true values. It is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \tag{24}$$

where N is the number of samples, y_i represents the predicted value, and $\hat{y}$ represents the true value. Corr is a metric used to measure the degree of similarity between the predicted results and the true labels. It is calculated using the Pearson correlation coefficient, which is defined as follows:

$$Corr = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{25}$$

where X represents the predicted results, and Y represents the true labels in our method.

4.5. Results

Table 1 shows the experimental results of the MCM method proposed in this paper on the MOSI and MOSEI datasets. We reproduced the best baseline Self-MM.

Table 1. The results on MOSI and MOSEI datasets. For Acc-2 and F1 score, the left of the "/" is classified as "negative/non-negative" and the right is classified as "negative/positive"; the same below. Models with * are reproduced under the same conditions.

Model	MOSI				MOSEI			
	Acc-2	**F1**	**MAE**	**Corr**	**Acc-2**	**F1**	**MAE**	**Corr**
TFN	-/80.8	-/80.7	0.901	0.698	-/82.5	-/82.1	0.593	0.700
LMF	-/82.5	-/82.4	0.917	0.695	-/82.0	-/82.1	0.623	0.677
MFN	77.4/-	77.3/-	0.965	0.632	76.0/-	76.0/-	-	-
RAVEN	78.0/-	76.6/-	0.915	0.691	79.1/-	79.5/-	0.614	0.662
MFM	-/81.7	-/81.6	0.877	0.706	-/84.4	-/84.3	0.568	0.717
MulT	81.5/84.1	80.6/83.9	0.861	0.711	-/82.5	-/82.3	0.58	0.703
Self-MM *	83.15/84.61	83.12/84.62	0.725	0.790	81.96/84.92	82.33/84.83	0.533	0.766
MCM	83.32/85.37	83.24/85.35	0.727	0.794	82.23/85.54	82.65/85.49	0.536	0.770

The results demonstrated that MCM outperformed all the baseline methods in terms of most evaluation metrics on both datasets, particularly in Acc-2 and the F1 score. Notably, MCM surpassed the performance of Self-MM, a method that leverages automatically generated labels for unimodal subtasks in multitask learning. Similar to Self-MM, previous multitask learning methods used in multimodal sentiment analysis have typically employed unimodal representations as subtasks, neglecting the learning of interactions between modalities. However, MCM addresses this limitation by jointly training single-mode and bimodal subtasks, considering both intramodal and intermodal interactions. This advancement in learning modal representations contributes to the overall improvement in model performance.

4.6. Ablation Study

To further analyze the contribution of each module in MCM, we conducted experiments on the MOSI dataset to compare the performance of models with different combinations of modules. The results of these experiments are shown in Table 2.

Table 2. Results of the ablation study on MOSI dataset. MT1 refers to the first group of subtasks in multitask learning, MT2 refers to the second group of subtasks in multitask learning, and CL refers to contrastive learning.

Model	Acc-2	F1 Score	MAE	Corr
Base	82.45/84.6	82.31/84.54	0.740	0.790
MT1	82.45/84.45	82.35/84.42	0.725	0.792
MT2	82.91/84.97	82.85/84.97	0.723	0.794
MT1/2	83.32/85.27	83.22/85.24	0.720	0.795
CL	82.62/84.63	82.52/84.6	0.731	0.790
MT1/2, CL(MCM)	83.32/85.37	83.24/85.35	0.727	0.794

The experimental results demonstrated that the complete MCM model outperformed the model without the multitask learning and contrastive learning modules across all evaluation metrics. Regarding multitask learning, the MCM model with both unimodal and bimodal subtasks achieved superior results compared to the model without any subtasks. This finding highlights the beneficial role of multitask learning in our model, as it enabled the model to leverage the shared information across different tasks and enhance overall performance.

The incorporation of contrastive learning led to better results compared to the model that solely included the multitask learning module. This outcome validated the effectiveness of contrastive learning in improving the multimodal sentiment analysis task. By

encouraging the consistency in the predictions, the contrastive learning module enhanced the reliability and robustness of the generated representations.

The inclusion of both unimodal and bimodal subtasks yielded superior performance compared to models with only one set of subtasks. This outcome suggested that the bimodal subtasks effectively constrained the learning of unimodal representations and prevented the model from overly focusing on intramodal interactions at the expense of neglecting intermodal interactions.

In addition to the ablation experiments conducted for different modules, we further analyzed the impact of each subtask in the multitask learning module. Specifically, we compared the performance of various combinations of unimodal or bimodal subtasks. By examining the results from these experiments, we gained insights into the individual contributions of each subtask in the multitask learning module of our model. These findings provided a deeper understanding of how the model benefited from the integration of different subtasks and shed light on the significance of capturing both intramodal and intermodal interactions for an effective multimodal sentiment analysis. The results of these experiments are shown in Table 3 for the unimodal subtasks and Table 4 for the bimodal subtasks, respectively.

Table 3. Results of the models containing different unimodal subtasks on the MOSI dataset.

Model	Acc-2	F1 Score	MAE	Corr
T	82.1/83.96	82.05/83.98	0.734	0.793
A	82.36/84.24	82.29/84.23	0.740	0.790
V	82.19/84.14	82.11/84.14	0.741	0.789
T, A	82.83/84.73	82.75/84.71	0.730	0.794
T, V	82.36/84.24	82.29/84.23	0.737	0.793
V, A	82.48/84.48	82.41/84.48	0.734	0.791
T, A, V	82.45/84.45	82.35/84.42	0.725	0.792

Table 4. Results of the models containing different bimodal subtasks on the MOSI dataset.

Model	Acc-2	F1 Score	MAE	Corr
AT	82.1/83.84	82.04/83.84	0.741	0.788
VT	82.74/84.7	82.69/84.7	0.734	0.793
AV	82.33/84.36	82.24/84.34	0.728	0.793
AT, VT	82.27/84.15	82.21/84.14	0.743	0.789
AT, AV	82.42/84.3	82.34/84.28	0.730	0.792
VT, AV	82.62/84.63	82.52/84.6	0.731	0.790
AT, VT, AV	82.91/84.97	82.85/84.97	0.723	0.794

Based on the results obtained from the two sets of experiments, it is observed that the performance of the model remained relatively similar when using only one or two subtasks. However, the improvement achieved in comparison to the base model was quite limited. This finding emphasized the necessity and effectiveness of joint training for both unimodal and bimodal representations in our model, which was consistent with our previous analysis regarding the purpose of incorporating unimodal and bimodal subtasks.

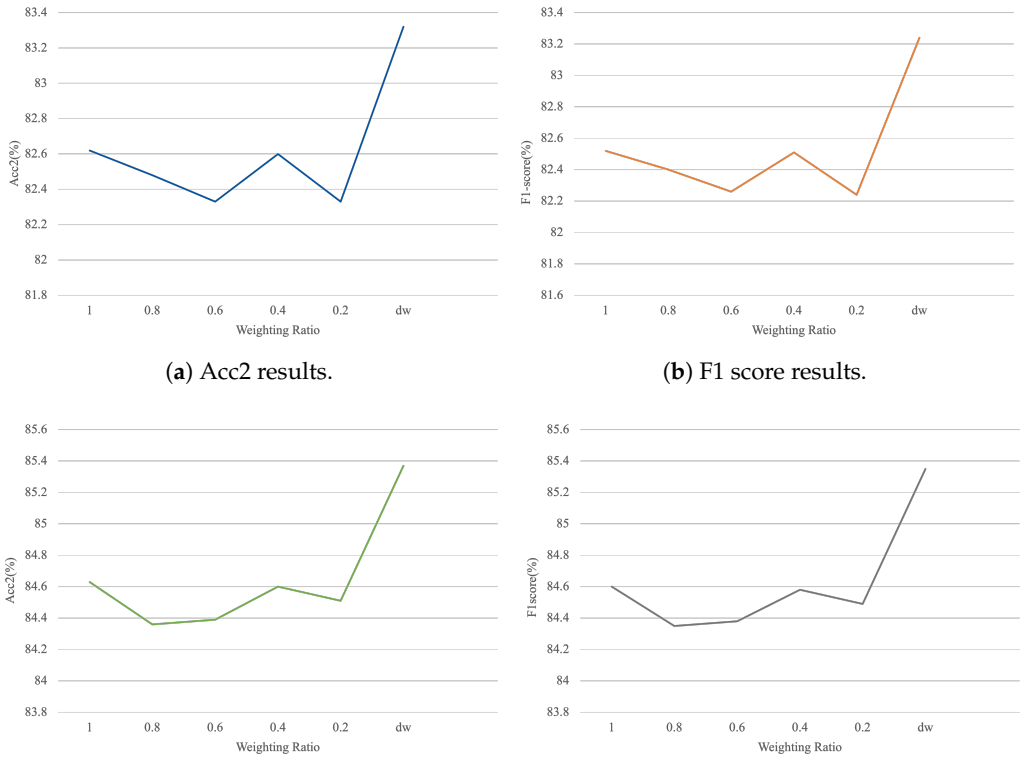
In summary, the experimental results stressed the effectiveness of multitask learning and contrastive learning in improving the multimodal sentiment analysis task. The combination of these modules, along with the inclusion of both unimodal and bimodal subtasks, led to significant performance improvements for the MCM model.

4.7. Dynamic Weights in Multitask Learning

In order to analyze the effect of dynamic weights in multitask learning in our model, we conducted five groups of comparison experiments. In these experiments, instead of adjusting the weights dynamically, we manually specified the weights of different tasks.

The weight w_m for the main task was set to a value in $\{1.0, 0.8, 0.6, 0.4, 0.2\}$, while the weight for the subtask was $(1 - w_m)$. The experimental results on the MOSI dataset are shown in Figure 6.

The results demonstrated that the dynamically adjusted weights method outperformed the manually adjusted weights method in terms of both Acc-2 and F1 score. This suggested that by using dynamic weights, we could obtain optimal weight configurations that led to improved prediction results. Additionally, the dynamic weights were learned simultaneously during model training, which offered the advantage of saving a significant time compared to manually adjusting weights.



(**a**) Acc2 results. (**b**) F1 score results.

(**c**) Acc2 results. (**d**) F1 score results.

Figure 6. The results on the MOSI dataset with different weights. The results in figure (**a**,**b**) are classified as "negative/non-negative". The results in figure (**c**,**d**) are classified as "negative/positive". The number in the independent variable indicates the weight of the main task, and *dw* indicates the dynamic weighting method in our model.

4.8. Case Study

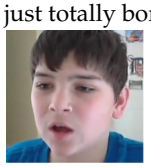
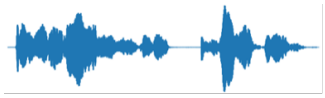
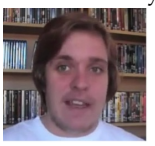
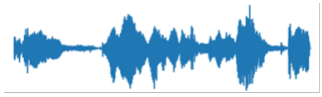
To assess the efficacy of learning modal representations with the assistance of multitask learning, we selected two samples from the MOSI dataset and analyzed their unimodal as well as fusion multimodal prediction results. For comparison, the results obtained from the model without the multitask learning module were also included. The experimental results are shown in Table 5.

To ensure the representativeness of the selected samples, we intentionally chose one sample with a positive sentiment polarity and another with a negative sentiment polarity. Analyzing the results presented in the table, we can observe that the multitask learning module had a more pronounced impact on the prediction results for the text modality compared to the video and audio modalities. This suggested that multitask learning could refine the text representations, which played a crucial role in the final prediction.

Furthermore, in the absence of the multitask learning module, the individual modal representations learned independently may contain misinformation that contradicts the true sentiment polarity. However, with the incorporation of multitask learning, these

misinformation effects could be effectively mitigated. By jointly training the model on multiple tasks, the conflicting information from individual modalities could be rectified, ensuring that the overall prediction remained consistent with the true sentiment polarity.

Table 5. Results of two selected samples from the MOSI dataset. The number before the text represents the true label of that sample. Positive sentiment polarity is represented by scores highlighted in green, where darker shades indicate a higher degree of positivity. Negative sentiment polarity is represented by scores highlighted in red, with darker shades indicating a higher degree of negativity.

Example	With Multitask	Without Multitask
(−2.4000) But when I saw it was just totally boring.  	M: −2.0249 T: −1.9917 A: −0.2503 V: −0.1818	M: −1.9375 T: −0.0810 A: −0.1243 V: 0.0928
(1.4000) Because um you really wanna see it for yourself.  	M: 1.4141 T: 1.5259 A: 0.5150 V: 1.0769	M: 1.3603 T: −0.0492 A: −0.1243 V: 0.0935

5. Discussion

With the development of multimedia, multimodal tasks have attracted increasing attention. Multimodal sentiment analysis aims to predict the sentiment polarity of data by integrating emotional information from multiple modalities. Current research primarily focuses on modal fusion, neglecting the importance of modality representation learning in prediction tasks. Therefore, we employed a multitask learning approach to assist in learning modality representation. In contrast to previous multitask learning methods, we designed subtasks specifically for bimodal representations generated by a gating mechanism. This allowed us to fully leverage the advantages of multimodal data, consider both intramodal and intermodal interactions, and enhance the model's ability to capture hidden emotional information in multimodal data. Additionally, we introduced a dynamic weight computation method to improve the performance of multitask learning. Furthermore, considering the presence of dropout layers in the model with the issue of inconsistencies between training and inference, we proposed a contrastive learning approach, which promoted consistency among the outputs of submodels with different dropout configurations, thereby strengthening the model's robustness and enhancing its performance. Through the auxiliary learning of modality representation and the resolution of dropout-related issues, this paper effectively improved the model's performance and holds practical value in real-world applications.

6. Conclusions

In this paper, we proposed MCM, a model that utilized multitask learning and contrastive learning to facilitate the learning of modal representations. A large number of previous studies of multimodal sentiment analysis take the modal representations obtained by training with a traditional neural network as the input to the modal fusion phase directly and focus their research on the fusion of multiple modal representations. However, modal data contain valuable emotional information that can significantly enhance the predictive power of the model. Therefore, we introduced a dynamic weighted multitask learning module to enable our model to capture the hidden information in the modal data. In addition, to alleviate the problem of inconsistent training and inference caused by the dropout layer, a contrastive learning module was added to further improve the effectiveness of our method. Experiment results indicated the efficacy of our proposed method.

During the experiment, it was observed that the prediction accuracy of audio, video, and video–audio representations was much lower than that of other modalities. This indicated that there was still room for improvement in the preprocessing of audio and video source data, as well as the fusion of audio and video features. In future research, we will continue to investigate and improve these areas.

Author Contributions: Conceptualization, M.Z. and X.W.; methodology, M.Z. and X.W.; software, M.Z.; validation, M.Z., X.W., Y.S. and B.C.; formal analysis, D.W.; investigation, Y.S.; resources, M.Z.; data curation, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, B.C.; visualization, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: MOSI and MOSEI can be downloaded at https://github.com/A2Zadeh/CMU-MultimodalSDK, accessed on 1 July 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.
2. Poria, S.; Cambria, E.; Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2539–2544.
3. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
4. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
5. D'mello, S.K.; Kory, J. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv. (CSUR)* **2015**, *47*, 1–36. [CrossRef]
6. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
7. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
8. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. *arXiv* **2022**, arXiv:2211.11256.
9. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning factorized multimodal representations. *arXiv* **2018**, arXiv:1806.06176.
10. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8992–8999.
11. Sun, L.; Lian, Z.; Liu, B.; Tao, J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Trans. Affect. Comput.* **2023**, 1–17. [CrossRef]
12. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 10790–10797.

13. Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; Liu, T.Y. R-drop: Regularized dropout for neural networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 10890–10905.
14. Ma, X.; Gao, Y.; Hu, Z.; Yu, Y.; Deng, Y.; Hovy, E. Dropout with expectation-linear regularization. *arXiv* **2016**, arXiv:1609.08017.
15. Zolna, K.; Arpit, D.; Suhubdy, D.; Bengio, Y. Fraternal dropout. *arXiv* **2017**, arXiv:1711.00066.
16. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.
17. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018; Volume 32.
18. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Conference. Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; NIH Public Access: Bethesda, MD, USA, 2019; Volume 2019, p. 6558.
19. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Virtual Event/Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
20. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the Conference. Association for Computational Linguistics, Online, 5–10 July 2020; NIH Public Access: Bethesda, MD, USA, 2020; Volume 2020, p. 2359.
21. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
22. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
23. Han, W.; Chen, H.; Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv* **2021**, arXiv:2109.00412.
24. Xue, X.; Zhang, C.; Niu, Z.; Wu, X. Multi-level attention map network for multimodal sentiment analysis. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 5105–5118. [CrossRef]
25. Cheng, H.; Yang, Z.; Zhang, X.; Yang, Y. Multimodal Sentiment Analysis Based on Attentional Temporal Convolutional Network and Multi-layer Feature Fusion. *IEEE Trans. Affect. Comput.* **2023**, 1–15. [CrossRef]
26. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 29–31 January 2019; Volume 33, pp. 7216–7223.
27. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [CrossRef]
28. Caruana, R. Multitask learning: A knowledge-based source of inductive bias1. In Proceedings of the Tenth International Conference on Machine Learning, Citeseer, Amherst, MA, USA, 27–29 June 1993; pp. 41–48.
29. Duong, L.; Cohn, T.; Bird, S.; Cook, P. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 845–850.
30. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1735–1742.
31. Gao, T.; Yao, X.; Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv* **2021**, arXiv:2104.08821.
32. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
33. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
34. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated multimodal units for information fusion. *arXiv* **2017**, arXiv:1702.01992.
35. Der Kiureghian, A.; Ditlevsen, O. Aleatory or epistemic? Does it matter? *Struct. Saf.* **2009**, *31*, 105–112. [CrossRef]
36. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5580–5590.
37. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 20–22 June 2016; pp. 1050–1059.
38. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv* **2016**, arXiv:1606.06259.
39. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
40. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 960–964.

41. Wood, E.; Baltrusaitis, T.; Zhang, X.; Sugano, Y.; Robinson, P.; Bulling, A. Rendering of eyes for eye-shape registration and gaze estimation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3756–3764.
42. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Continuous conditional neural fields for structured regression. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part IV 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 593–608.
43. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–3 February 2018; Volume 32.