applied sciences

MDPI

Article

Bimodal Fusion Network with Multi-Head Attention for Multimodal Sentiment Analysis

Rui Zhang [1,2], Chengrong Xue [1,2], Qingfu Qi [3], Liyuan Lin [2,*], Jing Zhang [1,2] and Lun Zhang [1,2]

[1] School of Software and Communications, Tianjin Sino-German University of Applied Sciences, Tianjin 300222, China
[2] College of Electronic Information and Automation, Tianjin University of Science & Technology, Tianjin 300222, China
[3] Gaussian Robotics Pte. Ltd., Tianjin 200100, China
* Correspondence: linly@tust.edu.cn; Tel.: +86-139-2069-0387

Abstract: The enrichment of social media expression makes multimodal sentiment analysis a research hotspot. However, modality heterogeneity brings great difficulties to effective cross-modal fusion, especially the modality alignment problem and the uncontrolled vector offset during fusion. In this paper, we propose a bimodal multi-head attention network (BMAN) based on text and audio, which adaptively captures the intramodal utterance features and complex intermodal alignment relationships. Specifically, we first set two independent unimodal encoders to extract the semantic features within each modality. Considering that different modalities deserve different weights, we further built a joint decoder to fuse the audio information into the text representation, based on learnable weights to avoid an unreasonable vector offset. The obtained cross-modal representation is used to improve the sentiment prediction performance. Experiments on both the aligned and unaligned CMU-MOSEI datasets show that our model achieves better performance than multiple baselines, and it has outstanding advantages in solving the problem of cross-modal alignment.

Keywords: multimodal sentiment analysis; bimodal fusion; multi-head attention

Citation: Zhang, R.; Xue, C.; Qi, Q.; Lin, L.; Zhang, J.; Zhang, L. Bimodal Fusion Network with Multi-Head Attention for Multimodal Sentiment Analysis. *Appl. Sci.* **2023**, *13*, 1915. https://doi.org/10.3390/app13031915

Academic Editor: Luis Javier Garcia Villalba

Received: 2 January 2023
Revised: 27 January 2023
Accepted: 29 January 2023
Published: 2 February 2023

1. Introduction

In recent years, multimedia forms such as audio and video have become important methods for people to obtain information. Text sentiment analysis (SA) has been extended to multiple scenes such as audio and video. Therefore, multimodal sentiment analysis has become a popular research direction [1–3], and its purpose is to effectively extract the sentiment intention of humans in a certain scene. By adopting various deep-learning techniques, many models have been developed for sentiment analysis tasks and have proven to be effective.

The existing sentiment analysis tasks are divided into single-modal sentiment analysis and multimodal sentiment analysis [4]. Traditional sentiment analysis tasks use single-peak text input and predict the sentiment attributes of the text content. Zhou et al. [5] proposed a text sentiment classification model using double-word embedding methods, which combines two models to represent the text to form a combinatory input of Bi-CNN. However, there are many multimodal (text, audio and video modalities) scenarios in real life, such as playing music with lyrics and video with subtitles. Moreover, multimodal inputs tend to convey more information than single-modal inputs, making it easy to make misjudgments in the single-modal case. For example, people can express opposite intentions by using different tones, speeds and volumes, and express their sentiments through facial expressions, body movements, situations and other information. Most of the existing multimodal sentiment analyses use three modalities and achieve good results. Xiao et al. [6] proposed a multichannel attentive graph convolutional network with cross-modality interactive learning and sentimental feature fusion. Mai et al. [7] considered sentiment intensity

attention and time-step level fusion and proposed a multiview sequential learning model to address the utterance-level human sentiment-comprehension problem. However, the researchers ignored situations where the image information was difficult to access, such as short videos with ruined images and TOEFL listening tests. In addition, image information contains more redundancy, which creates other interference factors to multimodal sentiment analysis.

Text and audio can complement each other well [8]. When there is ambiguity in the text, audio can obtain the sentiment information of the speaker. While it is difficult to obtain semantic information from audio, text can be supplemented. Some studies [9,10] show that the information provided by the text modality plays a leading role in multimodal sentiment analysis accuracy. Text can provide semantic information directly and effectively, which is supplemented by other modalities. However, the problem of "vector offset" will occur in the text modality superimposed with other modal information, which leads to the difference between the vector features obtained and the real sentimental features expressed by the text and finally affects the SA accuracy. Specifically, words that express sentiment in the text modality have a specific vector space, and the introduction of other modal information causes the intensity and direction of the word's vector to move in the original vector space.

Compared with single-modal sentiment analysis, bimodal sentiment analysis in the text and audio modalities also needs to consider the heterogeneity between different modalities, which will greatly increase the difficulty of the sentiment analysis. The sentiment expressed by text and audio do not correspond to each other, and different phonetic sentiments have different effects on the semantic meaning of the text. That is, the text and audio modalities often exhibit a "misaligned" nature. Some models [11–13] commonly implemented forced word alignment before training to solve the problem of "unaligned" nature, which aligned the visual and acoustic features to the resolution of words before inputting them into the model. The researchers [14–16] found that attention can capture the alignment between different modalities very well and achieve good results. However, there is no simple correspondence between the text modality and other modalities. A textual word can correspond to multiple frames of audio and images, and a single attention mechanism can only obtain a thin alignment.

To solve the above problems, we propose a bimodal fusion network with multi-head attention, which is a model for dealing with "misaligned" and "vector offset" multimodal languages. We illustrate the difference between the word alignment and the cross-modal attention inferred by our model in Figure 1. The main contributions of our paper are as follows:

- Provision of a novel model for processing text voice bimodal data that can solve the dynamic change problem of cross-modal data in the time dimension and update the cross-modal weight value by iteration;
- Solving the problem of long-term dependencies in intermodality and intramodality, focusing on solving the "vector offset" problem caused by audio modal data to the vector representation of textual words.

To verify the performance of our model in text–audio sentiment analysis, we conducted experiments on a standard CMU-MOSEI dataset. Experimental results show that our model achieves good results for text–audio sentiment analysis tasks.

This paper is organized as follows. Section 2 provides a brief literature review on modal sentiment analysis. In Section 3, we start by describing the proposed model in this paper, then introduce the proposed technique in detail. The experimental setup for evaluating the system and a discussion of the results achieved by various systems are presented in Section 4. Finally, the paper is concluded in Section 5.

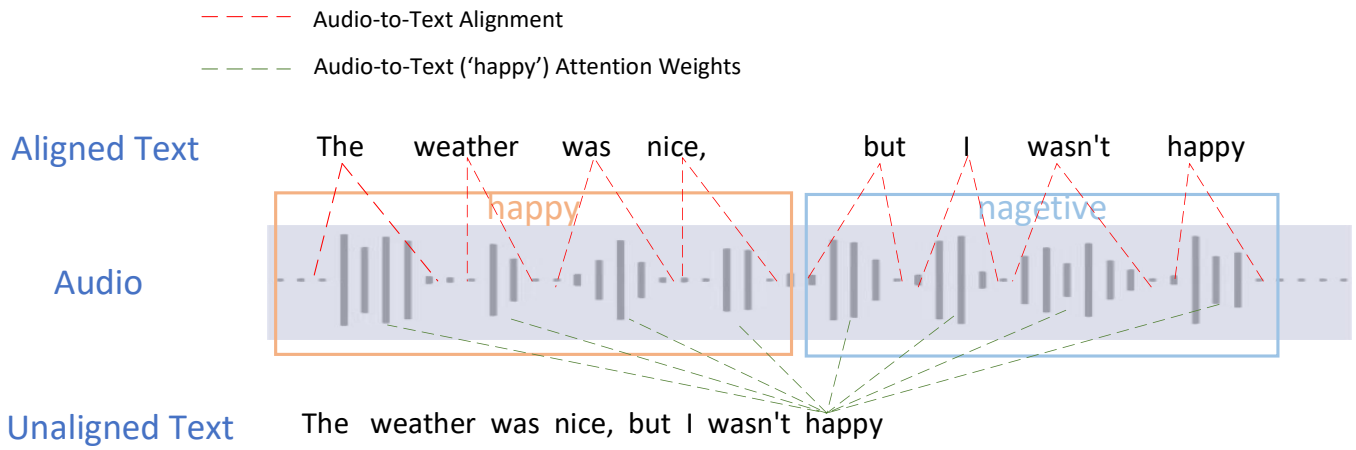

Figure 1. Example of word alignment by cross-modal attention.

2. Related Work

Multimodal sentiment analysis involves three elements: text, audio and video. Here, we introduce sentiment analysis tasks under text modality, audio modality and text–audio bimodality. The model that uses only one of text, acoustic or visual information is called a unimodal model, while the model that uses two or more kinds of information is called a multimodal model.

2.1. Unimodal Sentiment Analysis

2.1.1. Textual Sentiment Analysis (TSA)

The TSA method usually relies on "feature engineering" to obtain useful features related to sentiment, which can be divided into methods based on machine learning (ML) and depth learning (DL).

ML-based models generally rely on knowledge or statistical methods. The former models synonyms of large sentiment vocabulary, while the latter uses sentiment tags to mark available data in the database. Blekanov et al. [17] used a method based on multilingual knowledge to conduct sentiment analysis on events on the Twitter platform. Due to the limitation of knowledge, such methods can only be used to understand strictly defined concepts. TSA, which relies on statistical methods, overcomes the above problems and can process a large quantity of data. Pak et al. [18] proposed a subgraph-based model that represents documents as a set of subgraphs and inputs the features of these subgraphs into the SVM classifier. Some studies have used the two methods simultaneously and achieved good results [19].

DL-based models can automatically learn the feature representation used to identify emotions from data, including CNN and RNN. These methods are used to process TSA tasks at the document level [20], sentence level [21] and aspect (or word) level [22]. Yin et al. [21] made full use of word information and proposed a sentence-level sentiment analysis method using semantic words to enhance CNN. The authors of [22] proposed a graph-convolution network using SenticNet to obtain sentiment knowledge to enhance the sentence dependency, which realized efficient TSA.

2.1.2. Audio Sentiment Analysis (ASA)

The most common ASA method is to segment each audio into overlapping or nonoverlapping clips that are considered static and then extract audio features from the clips [1]. ASA can also be divided into methods based on ML and methods based on DL.

The early audio-feature extraction methods usually used machine learning to represent acoustic features. With the continuous involvement of psychology in machine learning, people have found that psychological research related to sentiment is of great help in extracting audio features. Therefore, people began to study the influence of sound parame-

ters on sentiment analysis, especially pitch, speed, intensity, duration and sound quality. Further research shows that other features [23] also play an important role in audio emotion analysis, including formant, pause, mel frequency cepstrum coefficients (MFCC), features based on energy operation, logarithmic frequency power coefficient (LFPC) and linear prediction cepstrum coefficient (LPCC).

Deep learning has received extensive attention in audio feature extraction. The authors of [24] used CNN to train the features extracted from all time frames, thus realizing the audio-emotion analysis task. However, such models cannot model the temporal information. To overcome this problem, long- and short-term memory (LSTM) is used to manually extract acoustic features [25]. The method based on deep learning does not consider feature engineering.

2.2. Text–Audio Bimodal Sentiment Analysis

The text–audio bimodal method uses both language content and voice cues to implement sentiment analysis tasks, and its performance is better than that of single-modality sentiment analysis tasks [26]. Cai et al. [27] proposed a multimodal emotion analysis model integrating CNN and LSTM, which can simultaneously capture spatial features and dynamic information. Pepino et al. [28] used BERT and openSMILE to obtain text features and acoustic features and adopted different fusion strategies to achieve sentiment analysis tasks. Later, people considered the alignment between different modalities. Xu et al. [15] used the attention mechanism to solve the alignment problem between voice frames and text words and achieved good results.

3. Proposed Approach

To solve the alignment problem of different modal information and alleviate the vector offset in multimodal representation learning, a bimodal multi-head attention network (BMAN) is proposed. The overall BMAN framework is shown in Figure 2. It takes text and audio sequences as the input. The main body of the model consists of two important parts: unimodal encoder and bimodal decoder. Unimodal encoders realize information interaction within unimodal to extract the effective syntactic and semantic features from the given unimodal inputs and include a text encoder and audio encoder. The bimodal decoder takes the outputs of two unimodal encoders as inputs and jointly decodes them to fully fuse intermodal information, where the acoustic information is used as auxiliary information to correct and supplement text semantics. Finally, the fused text representation serves for multimodal sentiment prediction.

Since previous research [9] has proven the remarkable ability of multi-head attention in capturing relationships, we take the transformer with multi-head attention as the backbone network of unimodal representation learning and cross-modal alignment and design a complete set of encoding–decoding structures to mine the dependence of different modal information on the time dimension, expecting to fit the complex alignment relationship between them. In contrast, the previous experimental results [29] show that removing the textual modality causes a large decline compared with the removal of other modalities, which demonstrates that the text modality plays a major role in multimodal sentiment analysis. To this end, we conscientiously consider that the different modalities deserve different weights and designed an auxiliary update scheme in which we took the text with direct semantics as the main knowledge of the network and updated it using the auxiliary modality. This can not only control the vector offset, but also ensure that the vector semantics are rich and accurate.

3.1. Unimodal Encoders

We utilize two independent transformer encoders to construct the text encoder and audio encoder, capturing text information and audio information, respectively. Given the language input $I_l = [l_1, l_2, \cdots, l_N]$, where $l_n = [i_1, i_2, \cdots, i_{T_l}]$, N and T_l are the total sample number and word number of each sample, respectively. We first use the GloVe embedding

tool to obtain the d_l-dimensional textual embeddings $X_n = [x_1, x_2, \cdots, x_{T_l}] \in \mathbb{R}^{T_l \times d_l}$. Similarly, given the audio input $I_A = [A_1, A_2, \cdots, A_M]$, where $A_m = [j_1, j_2, \cdots, j_{T_a}]$, M and T_a are the sample number of audio and the frame number of each sample, respectively. Their initial embeddings are represented as $X_a = [a_1, a_2, \cdots, a_{T_a}] \in \mathbb{R}^{T_a \times d_a}$, where T_a and d_a are the frame number and acoustic embedding dimension, respectively.

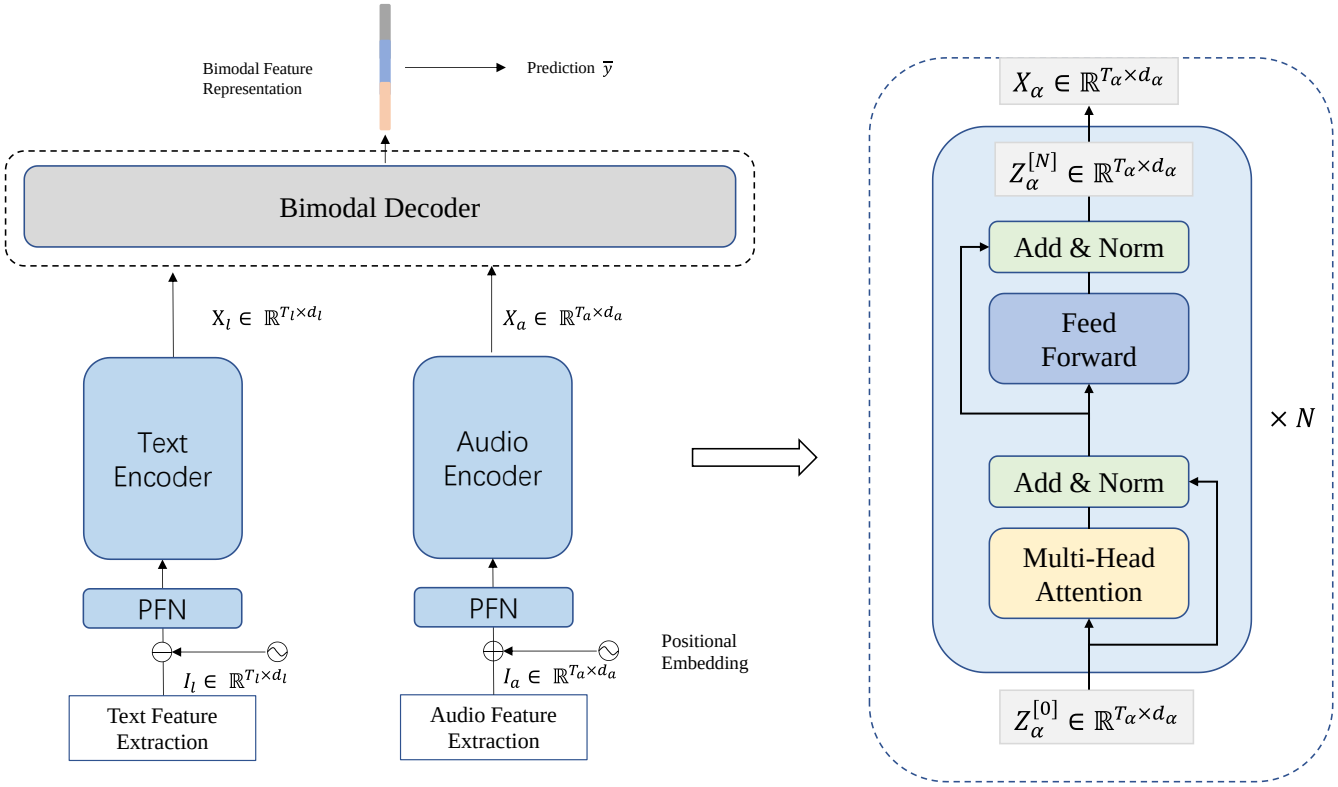

Figure 2. The overall architecture of BMAN. It includes two important components: unimodal encoders for capturing text and audio information and a bimodal decoder for jointly learning bimodal fused representation. The right part shows the detailed structure of the unimodal encoder, where Z can denote both textual and acoustic representations.

Taking the text encoder as an example, it extracts the relations of word pairs in parallel with the aid of attention and outputs the textual representation after full inter-word interaction. It consists of N stacked encoders, the core function layers of which are a multi-head attention sublayer and a feedforward sublayer, as shown in the right part of Figure 2.

The multi-head attention learns the token weights within the sequence in parallel, which indeed improves the computational efficiency but also means there is no order between words. However, natural language is a sequence of knowledge arranged in a certain order to express semantics, and this naturally determines the importance of order [30]. To make the model maintain efficient parallel learning while accounting for order information, we transform the absolute token position into position embedding (PE) following the previous work [31] and add it into textual embedding via a linear operation, called the position-aware feedforward network (PFN):

$$X_{PFN} = W(x_n + PE(x_n)) + b \tag{1}$$

X_l then serves as the input of the multi-head attention mechanism. The position information is updated along with the network and assists attention in weight assignment:

$$X^h = Softmax\left(\frac{QW_q^{(h)} \times \left(KW_k^{(h)}\right)^T}{\sqrt{d_l}}\right)V \tag{2}$$

where Q, K and V are equal to X_{PFN}, and the h denotes the head number. The matrix multiplication of Q and K obtains the interword correlation coefficient, and X^h is the representation after weighting. The necessity of the multi-head is to set up multiple semantic spaces to learn the complex relationship between words. We apply a linear operation on the concatenation of these heads to obtain the new representation:

$$X_{MHA} = concat\left(X^1, X^2, \cdots, X^H\right) \times W_o \tag{3}$$

Since we set N encoders, there are N multi-head attentions in the text encoder, leading to a deep network. Excessive layers may cause vanishing or exploding gradients, leaving the whole network unable to update. The residual connection adds the previous layer output to the current layer, preventing the vanishing gradient in backpropagation. Layer normalization can eliminate gradient explosion by limiting the output of each layer in a small numerical range. These two operations ensure the normal update of the network, and we apply them in every sublayer. Here, the sublayer means the multi-head attention sublayer:

$$X'_{MHA} = LayerNorm(X_{PFN} + X_{MHA}) \tag{4}$$

Since the above operations are all linear, a feedforward network with nonlinear activation is used to fit the complex semantics in the vector space:

$$X_{FFN} = \left(ReLU\left(X'_{MHA}W_1 + b_1\right)\right)W_2 + b_2 \tag{5}$$

where W_1, W_2, b_1, b_2 are trainable parameters of the linear layer and ReLU is the activation function. There should also be a residual connection and layer normalization after the feedforward sublayer:

$$X_l^{[0]} = LayerNorm\left(X'_{MHA} + X_{FFN}\right) \tag{6}$$

where $X_l^{[0]}$ denotes the textual representation obtained by the first encoder. As mentioned earlier, the text encoder consists of N encoders in a series. $X_l^{[0]}$ serves as the input of the next encoder, and the final output of the text encoder can be denoted as $X_l^{[N]}$.

Similarly, the audio encoder captures the interaction between acoustic information and obtains acoustic representation $X_a^{[N]}$.

3.2. Bimodal Decoder

As shown in Figure 3, the bimodal decoder takes the text encoder output X_l and audio encoder output X_a as input and conducts N joint decoding steps via N decoders in series to intelligently fuse the cross-modal information. Specifically, it matches the samples of each time step with that of another modality via multi-head attention. Although the alignment relationship between modalities is complicated, attention can adaptively allocate different weights to adapt to different alignment situations:

$$attention(Q, K, V) = softmax\left(\frac{QW_q^{(h)} \times \left(KW_k^{(h)}\right)^T}{\sqrt{d_k}}\right)V \tag{7}$$

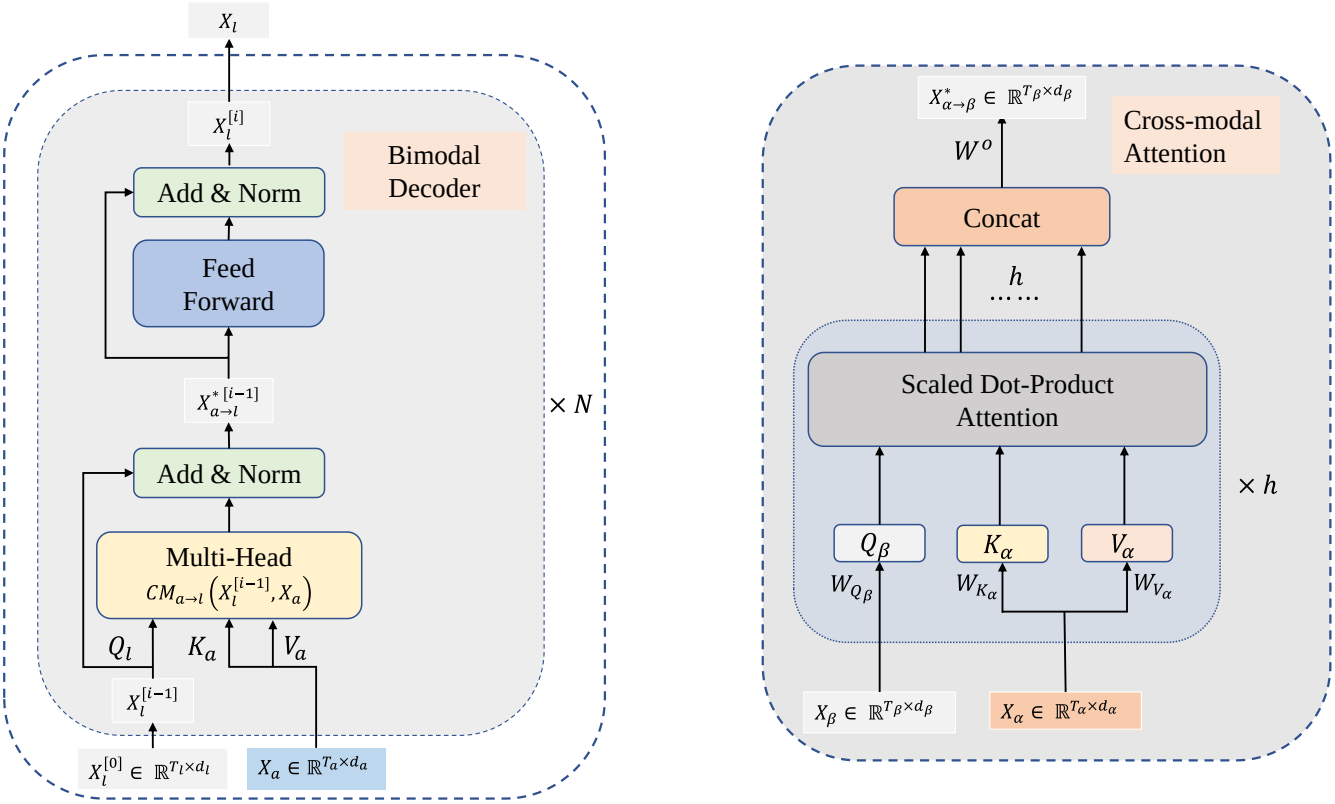

Figure 3. Overall architecture of the bimodal joint decoder (**left**) and detailed structure of the cross-modal attention mechanism (**right**).

We set the textual information X_l as query Q and the acoustic information X_a as K and V to assign weights to V according to Q. Equation (7) can be transformed as follows:

$$X_h^{[i]} = softmax\left(\frac{X_l^{[i-1]} W_l^{(h)} \times \left(X_a W_a^{(h)}\right)^T}{\sqrt{d_a}}\right) X_a \tag{8}$$

$$X_{a\to l}^{*[i]} = concat\left(X_1^{[i]}, X_2^{[i]}, \cdots, X_H^{[i]}\right) \times W_o \tag{9}$$

where $i \in [1, 2, \cdots, N]$ denotes the i-th decoder and $X_l^{[i-1]}$ is the output of i-1-th encoder. $W_l^{(h)}$ and $W_a^{(h)}$ denote the trainable weight matrices of the h-th attention head, and W_o is used to fuse the outputs of the H heads. Equation (8) alleviates the alignment problem and vector offset problem simultaneously. The i-th encoder obtains two vector sequences, one is the text information representation of each sample $X_l^{[i]}(l \in [1, N])$, and the other is the acoustic information representation of each frame $X_a^{[i]}(a \in [1, M])$. Attention will focus on the weighted fusion of these two sequences. Specifically, take the text sequence as the key, query the acoustic sequence, find the $N \times M$ weight matrix and multiply it by the text vector sequence to find the integrated representation. In this process, the text information of each sample will focus on the acoustic information of several frames related to it with high scores, and other irrelevant frame information will not be paid attention to. This enables acoustic–text alignment. In addition, attention can intelligently learn the correlation between acoustic information and text information. It realizes acoustic-to-text integration according to weights, which is more learnable than the direct concatenation of text sequence and acoustic sequence, thereby alleviating the uncontrolled vector offset.

We take textual information as the main representation by adding the weighted result with the input textual vectors instead of acoustic vectors:

$$X_l^{[i]} = LayerNorm\left(X_l^{[i-1]} + X_{a\to l}^{*[i]}\right) \tag{10}$$

After decoding N times, we finally obtain the intermodal textual representation $X_l^{[N]}$, which serves as the input of the prediction layer to calculate the probabilities of sentiment categories:

$$\bar{y} = softmax\left(W_c X_l^{[N]} + b_c\right) \tag{11}$$

where W_c and b_c are trainable parameters for the prediction layer.

4. Experiments

4.1. Dataset

Experiments were conducted on the public dataset CMU-MOSEI [10], which contains 22,856 manually annotated video segments from 250 topics. Its sentiment labels are divided into six fine-grained categories and range from the most negative label -3 to the most positive label 3. We split the dataset according to the ratio of 7:2:1 into three parts: training set (16,326 samples), validation set (1871 samples) and test set (4659 samples).

4.2. Baselines

We compared BMAN with a variety of multimodal sentiment analysis models:

Graph-MFN [10]: The graph memory fusion network constructs dynamic graphs for unimodal, bimodal and trimodal representations and achieves cross-modal interactions via dynamic connections between vertices.

EF-LSTM [32]: The early fusion LSTM (EF-LSTM) utilizes three LSTMs to extract information from every single modality and achieves early cross-modal fusion by concatenating their representations at each time step. It is suitable for aligned data.

BBFN(VA) [33]: A variant of the bimodal fusion network (BBFN) designs a bimodal complementary layer based on acoustic and visual modalities. It utilizes BiGRU to capture the internal dependency of each modality and then sets two gates to control the information interactions during the cross-modal multi-head attention process.

CTC+EF-LSTM [34]: It is a combination of connectionist temporal classification (CTC) [31] and EF-LSTM, where the CTC used for alignment prediction enables EF-LSTM to handle the unaligned data.

CTC+RAVEN [35]: This is a combination of CTC and RAVEN that dynamically fuses nonverbal information into textual representations by multimodal attention gating.

We also choose three unimodal models as baselines, called Transformer-T, Transformer-A and Transformer-V [9]. L, A and V denote text, audio and vision, respectively. They used a transformer structure to extract the unimodal utterance representations for the sentiment polarity prediction.

4.3. Setup

We chose the hyperparameters according to the neg/nonneg proxy in the validation results. The batch size was 32. The learning rate of the text encoder and audio encoder was 0.0001, and their attention heads were 1 and 48, respectively. The learning rate of the joint decoder was 0.001, and its number of attention heads was 12. The dimension of the feedforward network was 320 dimensions.

We used five common matrices to evaluate all the models, including the fine-grained accuracy Acc-7, binary accuracy Acc-2, binary F1 score, average absolute error (MAE) and Pearson correlation (Corr). Corr denotes the correlation coefficient between the prediction results and manual annotations. Note that we reported neg/nonneg and neg/pos according to whether the neutral labels are considered. The former judges the samples with nonnegative prediction scores as positive, while the latter strictly considers samples

with prediction scores greater than 0 as positive. We used the −/− marker to distinguish these two types of indicators, where neg/nonneg was on the left side and neg/pos on the right side.

4.4. Multimodal Sentiment Analysis Results

We conducted experiments on unaligned and aligned CMU-SOSEI and report the results as shown in Table 1. The proposed BMAN model presented overall advantages over the baselines.

Table 1. Comparison results (%) of multimodal sentiment analysis models. Bold indicates the optimal value of the corresponding evaluation indicators.

Model	Acc-7	Acc-2	F1-Score	MAE	Corr	Data Setting
Transformer-T	46.5	−/77.4	−/**78.2**	0.653	0.631	Unaligned
Transformer-A	41.4	−/65.6	−/68.8	0.764	0.31	Unaligned
Transformer-V	43.5	−/66.4	−/69.3	0.759	0.343	Unaligned
CTC+EF-LSTM	41.7	65.3/−	76.0/−	0.799	0.265	Unaligned
CTC+RAVEN	45.5	−/75.4	−/75.7	0.664	0.599	Unaligned
BMAN	**48.12**	**79.29/78.95**	**78.06**/77.84	**0.6471**	**0.640**	Unaligned
Graph-MFN	45.0	76.9/−	77.0/−	0.71	0.54	Aligned
EF-LSTM	46.7	**79.1**/72.02	**79.89**/61.89	0.674	**0.704**	Aligned
BBFN(VA)	41.1	−/71.1	−/64.5	0.816	0.261	Aligned
BMAN	**46.84**	75.55/**78.5**	76.11/**78.32**	**0.656**	0.624	Aligned

Obviously, BMAN outperformed unimodal models. Compared with Transformer-T, Transformer-A and Transformer-V, BMAN achieved average improvements of 4.32% (Acc-7), 9.15% (Acc-2) and 5.74% (F1), respectively. This is because these unimodal models focus on capturing intermodal dependency but lack support from other modal information. Moreover, we noticed that the unimodal model using audio only presented a great performance decline compared with the model using text only. This is consistent with our previous assumption that text plays the most important role in various modalities and implies the rationality of taking the text as the main information and audio as auxiliary information.

Compared with RNN-based models, such as CTC+EF-LSTM and CTC+RAVEN, BMAN achieved average improvements of 4.52% (Acc-7), 8.6% (Acc-2) and 1.99% (F1) on the unaligned dataset. This is reasonable because CTC+EF-LSTM fuses multimodal information via direct concatenation, ignoring the importance among various modalities as well as the vector offset during the fusion process, thereby presenting disappointing performance. CTC+RAVEN improved the flexibility of multimodal information fusion by placing multimodal attention gating after unimodal LSTM encoders; however, thin attention cannot fit the complex cross-modal interaction relationship. In contrast, BMAN achieved full cross-modal interactions via deep multi-head attention and abstractly fused acoustic information into textual representation in the form of joint decoding, alleviating the vector offset caused by audio modality to text modality.

In addition, we also observed that the proposed BMAN on the unaligned dataset has greater advantages over that on the aligned dataset. On unaligned CMU-MOSEI, BMAN outperformed CTC-LSTM with a considerable improvement of 6.42% (Acc-7), 13.99% (Acc-2) and 2.06% (F1), while this advantage appeared to be reduced on the aligned dataset. This is ascribed to our joint decoder being better at solving the alignment problem of multiple modalities. Its cross-modal attention assigns weights to all the audio frames according to the text samples. In this way, the text information of a certain sample will focus on the acoustic information of several frames related to it with high scores, and other irrelevant frame information will not be paid attention to, thereby achieving alignment between text samples and audio frames. Compared with models using CTC to obtain alignment sequences, our model internalized the data alignment into an intermediate step

of representation learning to facilitate the learning of the overall model, thereby presenting strong competitiveness on unaligned datasets.

5. Conclusions

In this paper, we proposed a multi-head attention bimodal fusion network for text–audio sentiment analysis tasks. The proposed network uses the multi-head attention mechanism to capture the attention weight of different audio to textual words and realizes cross-modal information alignment. The dual-modal decoder alleviates the vector offset problem caused by audio modality to text modality. Experiments show the proposed BMAN has good performance in multimodal sentiment analysis tasks. However, the design idea of our model is to give priority to the text modality, supplemented by the audio modality. Special situations where acoustic or visual modes are dominant are not fully considered, such as audio and video with ambiguous and sparse text. One direction in the future will be to explore a method to adaptively determine the dominant information in multimodal input.

Author Contributions: Conceptualization, R.Z., Q.Q., C.X., J.Z. and L.Z.; data curation, C.X. and Q.Q.; formal analysis, C.X. and Q.Q.; methodology, Q.Q.; project administration, R.Z.; resources, R.Z.; software, Q.Q.; supervision, L.L.; validation, R.Z., Q.Q. and C.X.; visualization, Q.Q.; writing—original draft, Q.Q. and C.X.; writing—review and editing, R.Z., C.X., Q.Q. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Tianjin Intelligent Manufacturing Special Fund Project-Underwater Heterogeneous Node Communication and Positioning Integrated Ad Hoc Network System Research and Development, Project No. 20201207, Tianjin Sino-German University of Applied Sciences Technology Project Grant No. ZDKT2018-006 and Tianjin Sino-German University of Applied Sciences Video Image Intelligent Analysis and Processing Technology Innovation Team and Tianjin "131" Innovative Talent Team.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Previously reported CMU Multimodal View Sentiment and Mood Intensity (CMU-MOSEI) data were used to support this study and are available at https://github.com/A2Zadeh/CMU-MultimodalSDK (accessed on 19 August 2021). These prior studies (and datasets) are cited at relevant places within the text as references.

Acknowledgments: In this section, you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: Author Qingfu Qi was employed by the company Gaussian Robotics Pte. Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Truong, Q.-T.; Lauw, H.W. VistaNet: Visual Aspect Attention Network for Multimodal Sentiment Analysis. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 305–312. [CrossRef]
2. Yang, K.; Xu, H.; Gao, K. CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20), Seattle, WA, USA, 12–16 October 2020; pp. 521–528.
3. Tu, G.; Wen, J.; Liu, C.; Jiang, D.; Cambria, E. Context- and Sentiment-Aware Networks for Emotion Recognition in Conversation. *IEEE Trans. Artif. Intell.* **2022**, *3*, 699–708. [CrossRef]
4. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2007**, *37*, 98–125. [CrossRef]
5. Zhou, M.; Liu, D.; Zheng, Y.; Zhu, Q.; Guo, P. A text sentiment classification model using double word embedding methods. *Multimed. Tools Appl.* **2022**, *81*, 18993–19012. [CrossRef]
6. Xiao, L.; Wu, X.; Wu, W.; Yang, J.; He, L. Multi-Channel Attentive Graph Convolutional Network with Sentiment Fusion for Multimodal Sentiment Analysis. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4578–4582.

7. Mai, S.; Hu, H.; Xu, J.; Xing, S. Multi-Fusion Residual Memory Network for Multimodal Human Sentiment Comprehension. *IEEE Trans. Affect. Comput.* **2022**, *13*, 320–334. [CrossRef]
8. Huang, Z.; Liu, F.; Wu, X.; Ge, S.; Wang, H.; Fan, W.; Zou, Y. Audio-Oriented Multimodal Machine Comprehension via Dynamic Inter- and Intra-modality Attention. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 13098–13106. [CrossRef]
9. 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 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569.
10. Bagher Zadeh, A.; 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, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2236–2246.
11. Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; Marsic, I. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2225–2235.
12. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. Memory Fusion Network for Multi-view Sequential Learning. *arXiv* **2018**, arXiv:1802.00927. [CrossRef]
13. Tsai, Y.-H.H.; Liang, P.P.; Zadeh, A.; Morency, L.-P.; Salakhutdinov, R. Learning Factorized Multimodal Representations. In Proceedings of the International Conference on Representation Learning, Addis Ababa, Ethiopia, 25–29 April 2019.
14. Lee, Y.; Yoon, S.; Jung, K. Multimodal Speech Emotion Recognition using Cross Attention with Aligned Audio and Text. *Interspeech* **2020**, 2717–2721.
15. Xu, H.; Zhang, H.; Han, K.; Wang, Y.; Peng, Y.; Li, X. Learning Alignment for Multimodal Emotion Recognition from Speech. *arXiv* **2020**, arXiv:1909.05645.
16. Yoon, S.; Byun, S.; Dey, S.; Jung, K. Speech Emotion Recognition Using Multi-hop Attention Mechanism. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2822–2826.
17. Blekanov, I.; Kukarkin, M.; Maksimov, A.; Bodrunova, S. Sentiment Analysis for Ad Hoc Discussions Using Multilingual Knowledge-Based Approach. In Proceedings of the 3rd International Conference on Applications in Information Technology, Wakamatsu, Japan, 1–3 November 2018; pp. 117–121.
18. Pak, A.; Paroubek, P. Text Representation Using Dependency Tree Subgraphs for Sentiment Analysis. *Database Syst. Adanced Appl.* **2011**, *6637*, 323–332.
19. Le, T. A Hybrid Method for Text-Based Sentiment Analysis. In Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 5–7 December 2019; pp. 1392–1397.
20. Liu, F.; Zheng, L.; Zheng, J. HieNN-DWE: A hierarchical neural network with dynamic word embeddings for document level sentiment classification. *Neurocomputing* **2020**, *403*, 21–32. [CrossRef]
21. Yin, R.; Li, P.; Wang, B. Sentiment Lexical-Augmented Convolutional Neural Networks for Sentiment Analysis. In Proceedings of the 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 26–29 June 2017; pp. 630–635.
22. Liang, B.; Su, H.; Gui, L.; Cambria, E.; Xu, R. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowl. Based Syst.* **2022**, *235*, 107643. [CrossRef]
23. Bitouk, D.; Verma, R.; Nenkova, A. Class-level spectral features for emotion recognition. *Speech Commun.* **2010**, *52*, 613–625. [CrossRef] [PubMed]
24. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
25. Atmaja, B.T.; Akagi, M. Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. In Proceedings of the 2019 IEEE International Conference on Signals and Systems (ICSigSys), Bandung, Indonesia, 16–18 July 2019; pp. 40–44.
26. Sebastian, J.; Pierucci, P. Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. *Proc. Interspeech* **2019**, 51–55.
27. Cai, L.; Hu, Y.; Dong, J.; Zhou, S. Audio-Textual Emotion Recognition Based on Improved Neural Networks. *Math. Probl. Eng.* **2019**, *2019*, 2593036. [CrossRef]
28. Pepino, L.; Riera, P.; Ferrer, L.; Gravano, A. Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6484–6488.
29. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP, Online, 3 August 2021; pp. 4730–4738.
30. Sun, Z.; Sarma, P.K.; Sethares, W.A.; Liang, Y. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8992–8999. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

32. Williams, J.; Kleinegesse, S.; Comanescu, R.; Radu, O. Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In Proceedings of the grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), Melbourne, Australia, 10 July 2018; pp. 11–19.
33. Han, W.; Chen, H.; Gelbukh, A.F.; Zadeh, A.; Morency, L.; Poria, S. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montréal, QC, Canada, 18–22 October 2021; pp. 6–15.
34. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
35. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 7216–7223. [CrossRef]