AMOA: Global Acoustic Feature Enhanced Modal-Order-Aware Network for Multimodal Sentiment Analysis

Ziming Li[1,2], **Yan Zhou**[1,*], **Weibo Zhang**[3], **Yaxin Liu**[1,2],
Chuanpeng Yang[1,2], **Zheng Lian**[4], **Songlin Hu**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]China Electronics Standardization Institute
[4]Institute of Automation, Chinese Academy of Sciences
{liziming, zhouyan, liuyaxin, yangchuanpeng, husonglin}@iie.ac.cn
zhangweibo@cesi.cn
lianzheng2016@ia.ac.cn

Abstract

In recent years, multimodal sentiment analysis (MSA) has attracted more and more interest, which aims to predict the sentiment polarity expressed in a video. Existing methods typically 1) treat three modal features (textual, acoustic, visual) equally, without distinguishing the importance of different modalities; and 2) split the video into frames, leading to missing the global acoustic information. In this paper, we propose a global **A**coustic feature enhanced **M**odal-**O**rder-**A**ware network (AMOA) to address these problems. Firstly, a modal-order-aware network is designed to obtain the multimodal fusion feature. This network integrates the three modalities in a certain order, which makes the modality at the core position matter more. Then, we introduce the global acoustic feature of the whole video into our model. Since the global acoustic feature and multimodal fusion feature originally reside in their own spaces, contrastive learning is further employed to align them before concatenation. Experiments on two public datasets show that our model outperforms the state-of-the-art models. In addition, we also generalize our model to the sentiment with more complex semantics, such as sarcasm detection. Our model also achieves state-of-the-art performance on a widely used sarcasm dataset.

1 Introduction

Multimodal sentiment analysis (MSA) has attracted more and more attention in recent years. In many cases, we need to combine the textual, acoustic, and visual features to predict sentiment polarity. For example, the tone of a person's voice and the changing expression can both have an impact on sentiment polarity prediction.

In most previous works, each modality will go through the same process at the fusion, or in other words, the three modalities are treated equally (Hasan et al., 2021; Chauhan et al., 2020). However, for sentiment analysis, the textual modality is usually the core modality based on life experience and previous works (Tsai et al., 2019; Han et al., 2021a; Hasan et al., 2021), because the text contains the most basic semantic information. The acoustic feature also plays an important role: a speech with a rising tone is more likely to express positive sentiment. Finally, facial expressions also have impacts on sentiment, such as the rising range of the corners of the mouth and the size of the pupils. However, the information about sentiment in expression is not as rich as that in tone. In addition, visual information does not always correspond to text like acoustic information. In many situations, the change in the speaker's facial expression is quite subtle. Even more, the speaker is absent in some videos. These visual noises may bring confusion to the model. Therefore, we consider the order of modalities, i.e. textual-acoustic-visual (t-a-v) while integrating them.

For MSA, a video is usually divided into many frames, and each frame corresponds to a very short time period in the video. The local acoustic features extracted from every single frame interact with each other in the fusion process. However, this method loses the global acoustic information and cannot fully reflect the tone feature of the whole audio.

To address these challenges, we propose a global acoustic feature enhanced modal-order-aware network. Firstly, the **M**odal-**O**rder-**A**ware network (MOA) is designed to integrate the three modalities in a certain order, where there are two stages con-

* Corresponding author.

Proceedings of the 29th International Conference on Computational Linguistics, pages 7136–7146
October 12–17, 2022.

necting the core and the outer modalities. We put textual modality at the core, and then the acoustic modality is integrated in stage 1, and finally, the visual modality is integrated in stage 2. At each stage, we design **C**ross-**M**odal **T**ransformer (CMT) based on the Transformer encoder (Vaswani et al., 2017) to integrate new modal features. Through CMT, the modality added before can also provide information for the later processes. Consequently, the textual feature learning is continuously enhanced through two stages, while the noise impact brought by the visual modality added in stage 2 is reduced. Then, to preserve the global acoustic information, we use *openSMILE* (Eyben et al., 2010) to extract the **G**lobal **A**coustic **F**eature (GAF) of the video to enhance modal feature learning. Furthermore, GAF and the multimodal fusion feature originally reside in their own spaces, which brings challenges to the fusion or concatenation. Inspired by MOCO (He et al., 2020), we employ contrastive learning to align the two features before concatenating them. Because visual modality may bring more noise and the processing of the entire video needs more computational power, we don't employ global visual features in our model. The main contributions of our paper are as follows:

- We propose AMOA - a novel multimodal sentiment analysis model that can integrate the three modalities in a certain order. In the modal-order-aware network, CMT is designed to fuse the features of different modalities.
- We are the first to introduce the global acoustic feature into MSA, which aims to preserve the global acoustic information and enhance the learning of the overall video feature. Furthermore, contrastive learning is utilized to align them before concatenation.
- We conduct experiments on sentiment (CMU-MOSI and CMU-MOSEI) and sarcasm (MUStARD) datasets, and the results show the state-of-the-art performance of our model.

2 Related Work

Multimodal fusion has always been the most critical step in MSA. Early works directly concatenate unimodal features or use outer product (Zadeh et al., 2018). With the development of the neural networks (Russakovsky et al., 2015; Hochreiter and Schmidhuber, 1997) and attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017), more and more complex networks have been applied to MSA to integrate modalities.

(Tang et al., 2021) uses a translation-based model to supplement the missing modalities. (Liu et al., 2021) is based on quantum probability modeling and uses multi-task learning to predict sentiment polarity and detect sarcasm at the same time. (Rahman et al., 2020) proposes an attachment to pre-trained language models so that they can adapt to the task of multimodal sentiment analysis. (Han et al., 2021a) performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations. (Han et al., 2021b) proposes a novel framework to maximize the mutual information in unimodal input pairs and between the multimodal fusion result and unimodal input. (Colombo et al., 2021) proposes new objectives to measure the dependency between modalities. These models treat all three modalities equally and mostly design very complex modules to achieve better results, while our model integrates three modalities in a certain order to distinguish their contributions, simple but effective.

Contrastive learning (CL) is a widespread self-supervised learning method in recent years. MOCO (He et al., 2020) and SimCLR (Chen et al., 2020) have achieved good results with CL in computer vision. After that, CL is applied to text-image multimodal tasks, such as image-text retrieval and visual question answering (Li et al., 2021).

3 Model

In this section, we will describe in detail how our proposed model works. The overall architecture of our model (AMOA) is shown in Figure 1. Our model consists of three modules: the modal-order-aware network (MOA), global acoustic feature (GAF) extraction & contrastive learning module, and classification module. In MOA, the unimodal features are first encoded and then integrated in a certain order, thus generating the multimodal fusion feature (Section 3.1). However, the multimodal fusion feature obtained in MOA is composed of single frame features and insufficient to reflect the overall change of tone, which is important for expressing sentiment. Therefore, we further extract GAF to complement the complete acoustic features. To align the multimodal fusion feature and GAF, we introduce contrastive learning and add the contrastive loss and classification loss together to guide

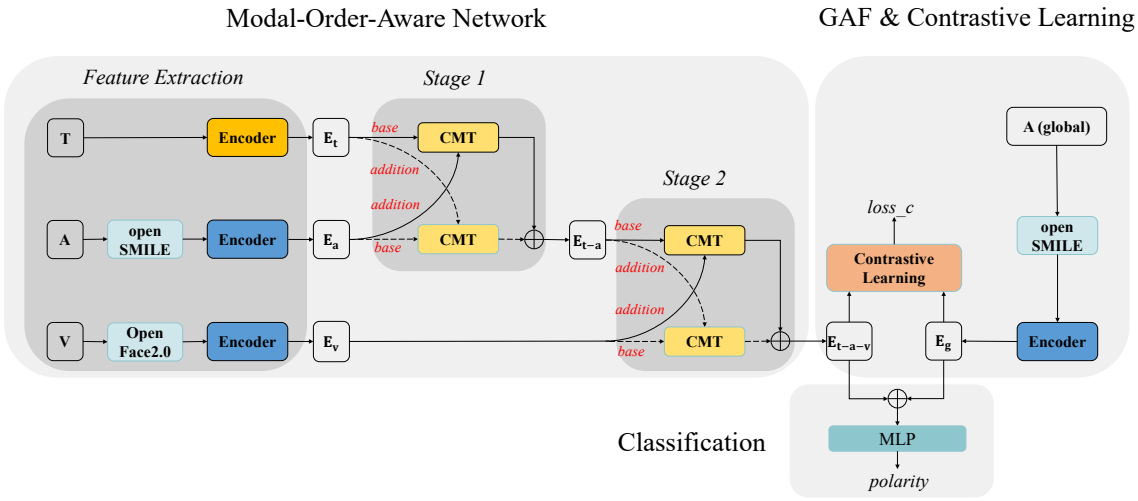


Figure 1: The overall architecture of our model. T, A, and V represent textual modality, acoustic modality (frame level), and visual modality (frame level) respectively. A (global) indicates acoustic modality which is not segmented. $\oplus$ represents the concatenation operation. The dashed parts make up two-way CMT.

model training. (Section 3.2) Finally, a multilayer perceptron layer is utilized for classification (Section 3.3).

3.1 Modal-Order-Aware Network

3.1.1 Unimodal Feature Extraction and Encoder

Textual: In this paper, pre-trained BERT-base-uncased (Devlin et al., 2019) is used as our text encoder, which has 12 layers and the hidden size is 768. The text encoder takes text $\mathbf{X}_t = \{x_1, x_2, \ldots, x_{n_t}\}$ in the video as input, and then output the last hidden layer representation: $\mathbf{E}_t \in \mathbb{R}^{n_t \times d_t}$, where n_t is the number of tokens and d_t is the hidden size (768) of BERT-base-uncased.

Acoustic: We use *openSMILE* to extract frame-level features of audio with 10 ms frame shift and 25 ms frame size. *openSMILE* provides a series of default feature sets, such as the INTERSPEECH 2010 Paralinguistic Challenge Feature Set (IS10) (Schuller et al., 2010), which contains different low-level features and their corresponding high-level features. For each frame, we extract IS10 as the feature vector: $\mathbf{X}_a = \{x_1, x_2, \ldots, x_{n_a}\}$, $\mathbf{X}_a \in \mathbb{R}^{n_a \times d_a}$ and n_a is the number of frames and d_a is the dimension of the acoustic feature. Then, we use P2FA (Yuan et al., 2008) to align the acoustic features to each word. Specifically, we obtain $\mathbf{X}_a^{'} = \{\mathbf{x}_1^{'}, \mathbf{x}_2^{'}, \ldots, \mathbf{x}_{n_t}^{'}\}$ by extracting the timing of all the words and averaging the acoustic feature vectors during this time. Because Transformer has the advantage of capturing long-distance dependencies, we directly use the Transformer encoder with random initialization to encode the feature and finally get the acoustic representation $\mathbf{E}_a \in \mathbb{R}^{n_t \times d_a}$.

Visual: The visual information in the video mainly comes from expressions, head shaking, and so on. We use OpenFace 2 (Baltrusaitis et al., 2018) to extract facial features at the frame level. These features are based on the Facial Action Coding System (Ekman and Rosenberg, 1997). Like acoustic modality, we then use the Transformer encoder to obtain visual representation $\mathbf{E}_v \in \mathbb{R}^{n_t \times d_v}$, where d_v is the hidden size of visual feature. The encoders of acoustic and visual modalities are independent of each other and they don't share any parameters.

3.1.2 Modal Order

Currently, most works treat the three modalities equally. They either feed the three modalities into a module for fusion and interaction at the same time, or integrate them in pairs. In this way not only the text information can not play a full role, but also the noise in the visual modality has the same impact. Therefore, we integrate three modalities in a certain order. First, the textual feature is extracted and encoded, i.e. $\mathbf{E}_t$; then, in stage 1, the acoustic features are fused to generate $\mathbf{E}_{t\text{-}a}$; finally, in stage 2, the visual features are integrated to generate $\mathbf{E}_{t\text{-}a\text{-}v}$. The final experiment will prove the optimality of the t-a-v order. At each stage, we employ CMT to integrate the previous and latter modalities, which will be described in detail in Section 3.1.3.

3.1.3 Cross-Modal Transformer

Based on the Transformer encoder, CMT is utilized to integrate new modalities, as shown in Figure 2. In stage 1, CMT inputs the textual and acoustic modal features and outputs the textual-acoustic fusion feature. Then in stage 2, the textual-acoustic feature and visual feature are fed to CMT and we obtain the textual-acoustic-visual fusion feature. Transformer is mainly composed of attention mechanism, and so is CMT. The input of the attention mechanism is a K-V pair and a query Q. The K-V pair can be regarded as basic information, while Q is additional information. We get the interactive information between Q and V by calculating the score between Q and K. Multi-head attention is composed of several parallel attention modules. They are individually responsible for calculating part of the results which will be concatenated into the final result. Different from the Transformer encoder, CMT utilizes multi-head attention instead of multi-head self-attention to integrate two features. Besides, CMT is also a multi-layer module. The original Transformer encoder cannot input two modalities and stack multiple layers together. Each layer of CMT has two inputs called *base* and *addition*. *base* of each layer stays the same, while *addition* is constantly updated.

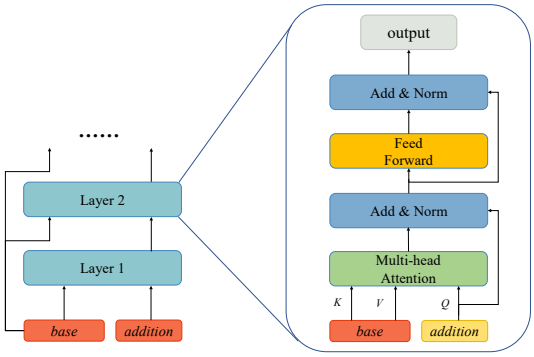


Figure 2: Cross-Modal Transformer.

Take the CMT in stage 1 as an example, in the first layer, *base* is the textual feature $\mathbf{E}_t$ and *addition* is the acoustic feature $\mathbf{E}_a$. Then we input *base* (K, V) and *addition* (Q) into the multi-head attention (MHA) followed by a residual connection and layer normalization:

$$\mathbf{Z} = MHA(addition,\ base,\ base), \quad (1)$$

$$\mathbf{Z}^{'} = Norm(\mathbf{Z} + addition). \quad (2)$$

Then we employ a feedforward neural network followed by a residual connection and layer normalization:

$$\mathbf{Z}^{''} = FeedForward(\mathbf{Z}^{'}), \quad (3)$$

$$\mathbf{E}_{t\text{-}a}^{(1)} = Norm(\mathbf{Z}^{'} + Dropout(\mathbf{Z}^{''})). \quad (4)$$

where $Norm$ is layer normalization. The above is the calculation process of the first layer, which can be expressed as:

$$\mathbf{E}_{t\text{-}a}^{(1)} = \mathrm{CMT}^{(1)}(\mathbf{E}_t,\ \mathbf{E}_a). \quad (5)$$

In the second and later layers, *base* is always the textual feature, and *addition* is the output of the previous layer:

$$\mathbf{E}_{t\text{-}a}^{(i)} = \mathrm{CMT}^{(i)}(\mathbf{E}_t,\ \mathbf{E}_{t\text{-}a}^{(i-1)}), i = 2, 3, \ldots, N, \quad (6)$$

where $\mathbf{E}_{t\text{-}a}^{(i)}$ is the output of the i^{th} layer of CMT and we take the output of the last layer as the textual-acoustic fusion feature $\mathbf{E}_{t\text{-}a}$.

Similar to the above procedure, we take $\mathbf{E}_{t\text{-}a}$ as *base* and the visual feature $\mathbf{E}_v$ as *addition* in CMT (*addition* of each layer is also updated continuously) in stage 2. Then we obtain the textual-acoustic-visual fusion feature:

$$\mathbf{E}_{t\text{-}a\text{-}v} = \mathrm{CMT}(\mathbf{E}_{t\text{-}a},\ \mathbf{E}_v). \quad (7)$$

In order to distinguish the importance of previous and latter modalities, we only use one-way CMT, that is, one input provides more information as *base* and another input provides less information as *addition*. For comparison, we also design a bidirectional module, as shown by the dashed line in Figure 1. The results from two CMTs are concatenated and transmitted to the next step. The model with one-way CMT is called S-AMOA and the model with two-way CMT is called B-AMOA. In this paper, we use S-AMOA by default.

Finally, a dropout layer and max-pooling layer are utilized to extract the most salient features across the time dimension:

$$\mathbf{E}_f = Maxpooling(Dropout(\mathbf{E}_{t\text{-}a\text{-}v})), \quad (8)$$

where $\mathbf{E}_f \in \mathbb{R}^{d_h}$ is the multimodal fusion feature.

3.2 GAF & Contrastive Learning

We use IS10 in *openSMILE* to extract GAF. Instead of splitting the video into frames, we extract the global acoustic feature of the whole video and get a one-dimensional feature vector: $\mathbf{X}_g \in \mathbb{R}^{d_g}$. We unsqueeze $\mathbf{X}_g$ to make it fit the input shape of

Transformer. Finally, we utilize the Transformer encoder to obtain the GAF representation: $\mathbf{E}_g \in \mathbb{R}^{d_g}$.

During training, apart from the final classification loss, we introduce another loss through contrastive learning, called $loss\_c$, to align the multimodal fusion feature $\mathbf{E}_f$ and global acoustic feature $\mathbf{E}_g$. One of the important steps of contrastive learning is to construct positive and negative samples. A direct idea is to take $\mathbf{E}_f$ and $\mathbf{E}_g$ belonging to the same sample in a batch as a positive pair and those not belonging to the same sample as negative pairs. Previous studies show that more negative samples promote contrastive learning. However, a larger batch size requires higher computing power so infinitely increasing the batch size is unpractical.

To applicably increase the number of negative samples, we construct a queue storing ($\mathbf{E}_f$, $\mathbf{E}_g$) pairs in the model. All the $\mathbf{E}_f$ in the queue are combined into the matrix $\mathbf{E}_f^q \in \mathbb{R}^{K \times d_h}$, and all the $\mathbf{E}_g$ in the queue are combined into the matrix $\mathbf{E}_g^q \in \mathbb{R}^{K \times d_h}$. K is the upper limit of the queue size. The data in the queue comes from the previous batches and acts as negative samples.

When a new batch comes, we get $\mathbf{E}_f^b \in \mathbb{R}^{B \times d_h}$ and $\mathbf{E}_g^b \in \mathbb{R}^{B \times d_g}$ which are positive samples of each other. B should be the batch size, but in practice, the utterance is processed with the context, so B is actually the product of the original batch size and the number of sentences in a sample. Next, we calculate the cosine similarity between each same row (i.e. the same sample) of the two matrices in this batch, which should be maximized:

$$\mathbf{S}_{pos} = Cosine(\mathbf{E}_f^b, \mathbf{E}_g^b), \tag{9}$$

where $Cosine$ is the cosine similarity function. Each value in $\mathbf{S}_{pos} \in \mathbb{R}^{B \times 1}$ is the similarity of the corresponding samples in this batch. Because there are multimodal fusion features and global acoustic features in both batch and queue, we construct double negative samples by calculating the similarity of $\mathbf{E}_f^b$ and $\mathbf{E}_g^q$ and the similarity of $\mathbf{E}_g^b$ and $\mathbf{E}_f^q$ respectively, which should be minimized:

$$\mathbf{S}_{neg}^{f \to g} = Cosine(\mathbf{E}_f^b, \mathbf{E}_g^q), \tag{10}$$

$$\mathbf{S}_{neg}^{g \to f} = Cosine(\mathbf{E}_g^b, \mathbf{E}_f^q), \tag{11}$$

where the value in the i^{th} row and the j^{th} column of $\mathbf{S}_{neg}^{f \to g} \in \mathbb{R}^{B \times K}$ is the cosine similarity between $\mathbf{E}_f$ of the i^{th} sample in the current batch and $\mathbf{E}_g$ of the j^{th} sample in the queue. Then we concatenate the three similarity matrices:

$$\mathbf{S} = Concat(\mathbf{S}_{pos}, \mathbf{S}_{neg}^{f \to g}, \mathbf{S}_{neg}^{g \to f}), \tag{12}$$

where the first column of $\mathbf{S} \in \mathbb{R}^{B \times (1+2 \times K)}$ is the similarity between positive samples, and the others are the similarity between positive samples and negative samples. Then we define a loss function to maximize the value of the first column of $\mathbf{S}$ and minimize the value of the other columns:

$$loss\_c = \frac{\sum_{i=1}^{B} | \, log(Softmax(\mathbf{S}_i)[0]) \, |}{B}. \tag{13}$$

As part of the final loss, $loss\_c$ will help to align $\mathbf{E}_f$ and $\mathbf{E}_g$.

Finally, we add $\mathbf{E}_f^b$ and $\mathbf{E}_g^b$ in the current batch to the queue, and if the queue size exceeds K, we pop up the pairs from the head of the queue.

3.3 Classification

We get the multimodal fusion feature $\mathbf{E}_f$ through MOA, and also extract the global acoustic feature $\mathbf{E}_g$. Now we concatenate the two features and then get:

$$\mathbf{R} = Concat(\mathbf{E}_f, \mathbf{E}_g), \tag{14}$$

where $\mathbf{R} \in \mathbb{R}^{(d_h+d_g) \times 1}$ is the final multimodal representation. Finally, We input the $\mathbf{R}$ into a multilayer perceptron (MLP) layer for classification:

$$\hat{y} = W_2(ReLU(W_1\mathbf{R})) + b_2, \tag{15}$$

where W_1, W_2, b_1 and b_2 are the parameters and $ReLU$ is the nonlinear activation function. During training, we use the MSE loss function to calculate the classification loss $loss\_f$ and then add contrastive learning loss with a certain weight:

$$loss = (1 - \alpha) \cdot loss\_f + \alpha \cdot loss\_c, \tag{16}$$

where α is a hyper-parameter, which is set to balance the two losses.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
MUStARD	552	69	69	690

Table 1: Split of three datasets.

Models	CMU-MOSI				CMU-MOSEI			
	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
TFN*	-/0.8080	-/0.8070	0.901	0.698	-/0.8250	-/0.8210	0.593	0.700
LMF*	-/0.8250	-/0.8240	0.917	0.695	-/0.8200	-/0.8210	0.623	0.677
MFM*	-/0.8170	-/0.8160	0.877	0.706	-/0.8440	-/0.8430	0.568	0.717
ICCN*	-/0.8300	-/0.8300	0.862	0.714	-/0.8420	-/0.8420	0.565	0.713
MulT*	0.8150/0.8410	0.8060/0.8390	0.861	0.711	-/0.8250	-/0.8230	0.580	0.703
MISA*	0.8079/0.8210	0.8077/0.8203	0.804	0.764	0.8259/0.8423	0.8267/0.8397	0.568	0.724
MAG-BERT*	0.8250/0.8430	0.8260/0.8430	0.731	0.789	0.8380/0.8520	0.8370/0.8510	0.539	0.753
self-MM*	0.8400/**0.8598**	0.8442/**0.8595**	**0.713**	**0.798**	0.8281/0.8517	0.8253/0.8530	0.530	0.765
MMIM‡	0.8324/0.8521	0.8311/0.8515	0.722	0.786	0.8418/0.8558	0.8425/0.8535	0.538	0.763
BBFN‡	0.8134/0.8353	0.8124/0.8351	0.833	0.743	0.8298/0.8569	0.8327/0.8570	0.579	0.759
S-AMOA (ours)	**0.8411**/0.8415	**0.8452**/0.8421	0.720	0.788	**0.8560/0.8645**	**0.8601/0.8654**	**0.526**	**0.772**
B-AMOA (ours)	0.8163/0.8277	0.8173/0.8283	0.735	0.786	0.8501/0.8575	0.8508/0.8587	0.578	0.766

Table 2: Performances of multimodal models on the CMU-MOSI and CMU-MOSEI datasets. * indicates that the results are from (Han et al., 2021b). ‡ indicates that the results are reproduced from open-source code with hyper-parameters provided in original papers. For Acc-2 and F1, we have two methods of calculation: non-negative/negative (left) and positive/negative (right). The best results are marked in bold. ↑ indicates that the higher the value, the better the result; ↓ indicates that the lower the value, the better the result. Bolded numbers represent the best results.

Models	Acc-2	F1
MFN◇	0.7391	0.7386
MulT◇	0.7536	0.7541
MAG(BERT)◇	0.7826	0.7818
MAG(XLNet)◇	0.7681	0.7679
A-MTL†	-	0.7657
QPM†	-	0.7753
HKT†	0.7941	0.7925
S-AMOA (ours)	**0.8406**	**0.8412**
B-AMOA (ours)	0.8116	0.8116

Table 3: Performances of multimodal models on the MUStARD dataset. ◇ are the results on the dataset using the original code provided in the paper. † indicates that the results are from the original paper. - indicates that the original paper provides neither the results under the Acc metric nor the training code.

4 Experiments

4.1 Datasets and Metrics

In order to verify the performance of our model in sentiment polarity prediction, we conduct experiments on two widely used public datasets: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh and Pu, 2018). CMU-MOSI has 2199 video clips, each of which is a speaker sharing their opinions on something. Each clip is labeled with the polarity of sentiment, and the range of labels is: [-3, 3]. CMU-MOSEI has 23454 film review clips, which are labeled in the same way as CMU-MOSI. Our model is not only applicable to the prediction of general sentiment polarity, but also can detect more complex sentiments, such as sarcasm. To verify this, we conduct experiments on the MUStARD dataset (Castro et al., 2019), the unique multimodal sarcasm dataset containing three modalities. The dataset is collected from four TV shows, with a total of 690 samples.

We use four commonly used evaluation metrics to evaluate the performance of the model on the MOSI and MOSEI datasets: binary classification accuracy (Acc-2), which divides seven labels into two categories (positive/negative and non-negative/negative); binary classification F1; mean absolute error (MAE), which is the difference between the predicted value and the real value; Pearson correlation (Corr), which measures the degree of prediction skew. The label of each sample is sarcasm or non-sarcasm, so we only use Acc-2 and F1 to evaluate the performance on the MUStARD dataset.

The split specifications of the three datasets are provided in Table 1. To motivate future research, the code will be released soon.

4.2 Baselines

For sentiment polarity prediction, we compare our model with many baseline models.

TFN (Zadeh et al., 2017): It integrates three modal features by outer product, which is a very classic work.

LMF (Liu et al., 2018): It performs multimodal fusion using low-rank tensors to improve

efficiency.

MFM (Tsai et al., 2018): It decomposes features into modal fusion features and modal specific features to enhance model robustness.

ICCN (Sun et al., 2020): It obtains multimodal embedding by calculating the outer product of the text and the other two modalities.

MulT (Tsai et al., 2019): It uses cross-modal transformers to fully integrate three modalities for aligned sequences or unaligned sequences.

MISA (Hazarika et al., 2020): It projects the modalities into two different subspaces to learn the intra modal features and inter modal features, respectively.

MAG (Rahman et al., 2020): It adds a multimodal adaptation gate to the existing pre-trained language models (BERT and XLNet) so that they can receive acoustic and visual information during fine-tuning. Because our model uses BERT for word embedding, MAG-BERT is employed as a baseline.

self-MM (Yu et al., 2021): It generates uni-modal labels based on self-supervised learning, and then jointly trains uni-modal and multi-modal tasks.

MMIM (Han et al., 2021b): It maximizes the mutual information in a multimodal fusion pipeline to maintain task-related information.

BBFN (Han et al., 2021a): It focuses on bimodal fusion process and balances the contribution of different modality pairs properly.

Furthermore, we also select some sarcasm detection baselines for comparison on the MUStARD dataset.

MFN (Zadeh et al., 2018): It obtains the intra-modal information and inter-modal information based on LSTM and passes the multimodal fusion information through time.

A-MTL (Chauhan et al., 2020): It manually annotates the samples in the MUStARD dataset with sentiment and emotion as well as analyzes sarcasm, sentiment, and emotion together through multi-task learning.

QPM (Liu et al., 2021): It builds a quantum probability-driven multi-task learning framework, including a quantum-like fusion network and quantum incompatibility measurements.

HKT (Hasan et al., 2021): Besides the three modalities, it introduces the ambiguity of words and sentiment dictionary and constructs a bimodal cross-attention layer based on Transformer.

4.3 Main Results

The experimental results on the MOSI and MOSEI datasets are shown in Table 2. On the MOSEI dataset, our model outperforms all baseline models in every metric. In the binary classification task, our model attains an improvement of **1**% - **2**% over other models, which indicates the advantage of our model in sentiment polarity prediction. On the MOSI dataset, our model outperforms all baseline models in Acc-2 and F1 (non-negative/negative). In other metrics, our model also achieves results close to SOTA. It is worth noting that the advantage of our model in Acc-2 (non-negative/negative) is more obvious than that in Acc-2 (negative/positive). This is because our model tends to classify the samples labeled neutral into the positive category, which is consistent with life experience.

The experimental results on the MUStARD dataset are shown in Table 3. The results illustrate that our model achieves the best performance and outperforms all baseline models (+**4.65**%). Some baseline models use context information (A-MTL, QPM, HKT), but the results are worse than our model without context information.

These results demonstrate the superiority of our proposed model and indicate the effectiveness of the modal-order-aware network and GAF compared with all baseline models.

4.4 Analysis

In order to further analyze the performance of our model and verify the contribution of each module, we conduct extensive experiments on the MOSI and MOSEI datasets.

4.4.1 CMT

We design two kinds of CMT, one-way and two-way, and their experimental results are shown in Table 2 and Table 3. The results show that the two-way CMT enhances the noise influence of the latter modality, and makes the previous modality unable to play a full role, which has an adverse impact on the performance of the model.

The number of layers N of CMT in the model is also a hyper-parameter. We set different N and conduct experiments on MOSI and MOSEI datasets. The results in Figure 3 show that when N is 5, the model achieves the best performance on the MOSI dataset, and when N is 2, the model achieves the best performance on the MOSEI dataset. As the number of layers increases, the information captured by CMT also increases. However, CMT may

Models	CMU-MOSI				CMU-MOSEI			
	Acc-2	F1	MAE	Corr	Acc-2	F1	MAE	Corr
AMOA	**0.8411/0.8415**	**0.8422/0.8421**	**0.720**	**0.788**	**0.8560/0.8645**	**0.8601/0.8654**	**0.526**	**0.772**
t-v-a	0.8265/0.8307	0.8287/0.8335	0.737	0.786	0.8521/0.8614	0.8555/0.8639	0.589	0.771
a-t-v	0.8279/0.8338	0.8291/0.8327	0.739	0.780	0.8542/0.8608	0.8576/0.8630	0.582	0.770
a-v-t	0.8236/0.8323	0.8255/0.8314	0.748	0.781	0.8499/0.8611	0.8542/0.8637	0.584	0.771
v-t-a	0.8250/0.8262	0.8255/0.8303	0.741	0.775	0.8499/0.8622	0.8530/0.8640	0.575	0.769
v-a-t	0.8309/0.8262	0.8314/0.8303	0.744	0.774	0.8527/0.8617	0.8565/0.8633	0.581	0.771
−v	0.8250/0.7988	0.8252/0.7978	0.733	0.770	0.8492/0.8564	0.8506/0.8573	0.578	0.759
−a	0.8090/0.8231	0.8091/0.8240	0.756	0.769	0.8475/0.8581	0.8472/0.8590	0.579	0.757
−t	0.7116/0.7215	0.7168/0.7015	0.899	0.335	0.7391/0.7095	0.7639/0.7252	0.795	0.392
−GAF	0.7804/0.8192	0.7791/0.8224	0.760	0.761	0.8486/0.8564	0.8498/0.8567	0.573	0.761
−CL	0.8265/0.8033	0.8290/0.8099	0.739	0.769	0.8518/0.8603	0.8537/0.8612	0.587	0.761
−CL (f→g)	0.8309/0.8368	0.8314/0.8387	0.731	0.771	0.8544/0.8608	0.8580/0.8623	0.583	0.769
−CL (g→f)	0.8294/0.8246	0.8291/0.8259	0.734	0.772	0.8555/0.8611	0.8595/0.8623	0.581	0.771
−q	0.8236/0.8105	0.8251/0.8112	0.733	0.771	0.8520/0.8625	0.8541/0.8645	0.600	0.762

Table 4: Order study and ablation study. $-m$ means to remove the m mode, where $m \in$ {t, a, v} is the three modalities. −GAF means not using GAF to enhance feature learning. −CL means to directly concatenate the multimodal fusion feature and GAF without contrastive learning for alignment. −CL (f→g) means to remove half of the negative samples calculated by Eq.(10) and −CL (g→f) means to remove half of the negative samples calculated by Eq.(11). −q means to construct positive and negative samples only from the same batch without using the queue.

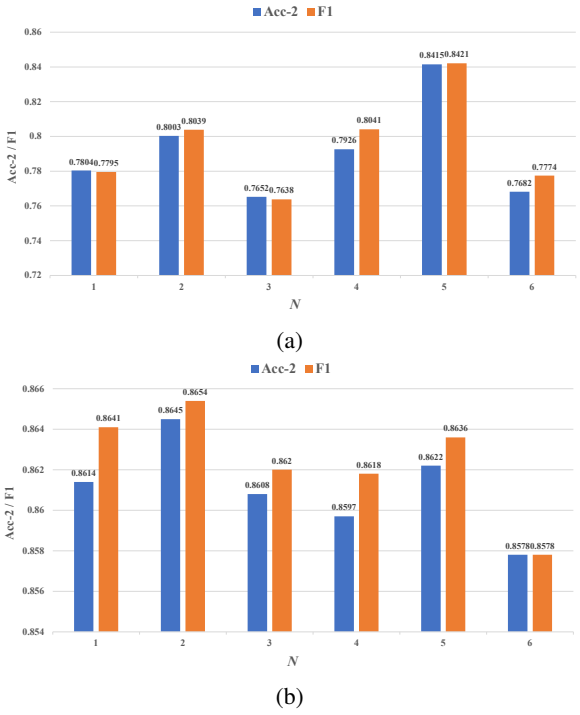


Figure 3: Experimental results with different layers of CMT. (a) is on the CMU-MOSI dataset. (b) is on the CMU-MOSEI dataset.

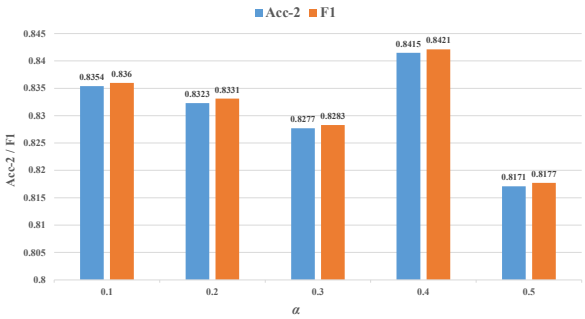
suffer from the distribution shift in the up layers, which will make the similarity less reliable. So bigger isn't necessarily better for N. Please refer to the experiment results in the supplementary materials.

4.4.2 Order of Modalities

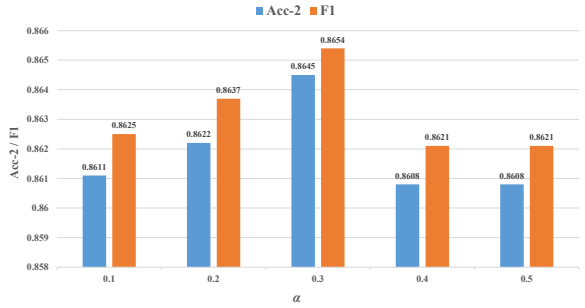
In our modal-order-aware network, we put text at the core and then integrate acoustic modality in stage 1, and visual modality is integrated in stage 2 (i.e., t-a-v). We try all permutations and obtain the prediction results under the same settings, as shown in Table 4. The order of t-a-v performs best, which verifies our hypothesis of modal order.

4.4.3 Role of Unimodalities

In order to verify the role of every single modality, we separately remove one modality and integrate the other two modalities in the original order. For example, the fusion order t-v removes the acoustic modality (−a). The experimental results are shown in Table 4. When a modality is removed, the performance of the model decreases in varying degrees, which shows that each modality plays an important role. Specifically, when the textual modality is removed, the performance decreases most obviously. In addition, The impact of acoustic modality is slightly greater than that of visual modality.



(a)



(b)

Figure 4: Experimental results with different α. (a) is on the CMU-MOSI dataset. (b) is on the CMU-MOSEI dataset.

4.4.4 Role of GAF

To verify the role of the global acoustic feature, we remove GAF and the contrastive learning module from our model. The results are shown in Table 4 ($-$GAF). Without GAF, model performance drops to some extent, which indicates that GAF plays an important role in our model. In addition, GAF is more effective in MOSI dataset. This is related to the different distribution of data in the two datasets. In many cases, GAF contributes more to the analysis of the examples which are relatively short but have large sound fluctuations. The proportion of this kind of example in MOSI is greater than that in MOSEI.

4.4.5 Role of Contrastive Learning

Furthermore, we remove the contrastive learning module from the original model and concatenate the multimodal fusion feature and GAF directly to verify the role of contrastive learning. The results in Table 4 ($-$CL) show that when the contrastive learning module is removed, we can see a clear drop in all metrics. In contrastive learning, we construct double negative samples based on different $\mathbf{E}_f$ and $\mathbf{E}_g$. When only one group negative samples are used, the performance of the model decreases to varying degrees.

In our experiments, we set the hyper-parameter α as the weight of the loss of the contrastive learning module in the whole loss. The influence of the value of α on the experimental results is shown in Figure 4. When α is 0.4, the model achieves the best performance on the MOSI dataset. When α is 0.3, the model achieves the best performance on the MOSEI dataset.

A queue is used to construct more negative samples. We try to remove this queue and construct negative samples only in the same batch. As shown in Table 4 ($-$q), the queue plays an important role in the contrastive learning module of our model.

5 Conclusion

For multimodal sentiment analysis, we propose the modal-order-aware network to integrate the three modalities in a certain order to distinguish the importance of different modalities. Besides, we are the first to introduce the global acoustic feature into this task to capture the changes in the tone of the whole video. Considering the misalignment between the multimodal fusion feature and GAF, contrastive learning is utilized to align them before concatenation. Experiments on three widely used datasets show that our model achieves the best performance. Besides, we also verify the effectiveness of each module of our model.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 4351–4360.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 231–245.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15.

Wei Han, Hui Chen, and Soujanya Poria. 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12972–12980.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

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

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. 2021. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 871–880.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256.

Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2020:2359–2369.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2794–2797.

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

Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong. 2021. Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5301–5311.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan

Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. In *International Conference on Learning Representations*.

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

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

Jiahong Yuan, Mark Liberman, et al. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

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

Amir Zadeh and Paul Pu. 2018. 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 (Long Papers)*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.