Modalities Missing in Multimodal Sentiment Analysis

Yisheng Zhu *

Houston International Institution, Dalian Maritime University, Dalian, 116000, China

* Corresponding Author Email: zys040509@dlmu.edu.cn

Abstract. In recent years, sentiment analysis has been widely applied in various fields such as smart home, intelligent healthcare, and education. Multiple technologies such as eye movement recognition and text analysis are employed as means to determine the emotional state of users. However, during this process, factors such as sensor failures, privacy restrictions, and environmental disturbances can all lead to the problem of modal absence in sentiment analysis. Traditional multimodal approaches often fail to handle the correlations between different modalities during feature extraction, and when integrating features, they often neglect multi-scale emotional cues. This research aims to summarize typical multimodal feature fusion models. Based on four optimized models developed in recent years, it aims to summarize and outline the technical breakthroughs in the emerging model architectures in addressing the issue of modality absence in multimodal sentiment analysis, and also makes assumptions and prospects for future multimodal sentiment analysis in dealing with modality absence.

Keywords: Multimodal sentiment analysis, Modal absence, Feature fusion, Model framework optimization.

1. Introduction

In recent years, multimodal sentiment analysis has become a research and application hotspot in many fields. Its aim is to precisely identify and understand human emotional states by integrating multiple modalities such as audio, text, and video [1]. The data shows that the SSC305DE product of Xingchen Technology collects data through the end-side device using dual cameras and voice sensors, and combines with the AI large model for cloud confirmation. It can identify the behavior of elderly people falling and their pain emotions, and actively notify their families. The application of this multimodal technology reduces the false alarm rate of fall event identification by 45% and the response time is less than 2 seconds. In the pilot communities, the number of elderly safety incidents decreased by 60%, and the satisfaction rate for emotional needs increased by 78%. However, due to factors such as environmental disturbances, privacy restrictions, and equipment failures, the absence of modalities has brought many inconveniences to multimodal emotion analysis, significantly affecting the analysis performance [2]. Due to the dim lighting at night, the in-vehicle emotion detection system is unable to monitor the driver's facial expressions. It mistakenly interprets the outbursts and sudden acceleration as "road rage", resulting in incorrect system prompts. Therefore, the significance of exploring the issue of modal absence in multimodal sentiment analysis is self-evident. However, traditional multimodal analysis often neglects the importance and correlation of features when dealing with modal absence problems, and when integrating, it often overlooks multi-scale emotional cues. This exploration will summarize typical multimodal feature fusion models, based on four optimized models, summarize and induce the technological breakthroughs of emerging model architectures in addressing the issue of missing modalities in multimodal sentiment analysis, and make assumptions and prospects for future multimodal sentiment analysis to address missing modalities.

2. Research methods and process

2.1. Typical methods

Table 1. Typical Architecture for Modal Missing Feature Fusion

Architecture type	Feature fusion mechanism	Architecture type	Feature fusion mechanism
Early fusion	Input layer concatenation, Average pooling	Modular design makes it easy to expand new modes	Performance deteriorates when modal differences are significant
Deep fusion	Cross modal attention, double linear fusion	Modal interaction fully supports arbitrary modal combination of input and output	Inference latency is high
Generative fusion	Generative adversarial networks	Good robustness under the high fault rate and enhancing the data generalization	The training is difficult to converge and unstable
Transformer	Multi-head self-attention, feedforward neural network, residual connection, cross-attention mechanism	excellent long-distance dependency capture capability, strong parallel computing capability	Modal gap, large parameter scale, high data cost

As shown in Table 1, simple concatenation and pooling arithmetic operations during early fusion can lead to the loss and dilution of modal features, making it impossible to adapt to the significant differences in the distribution of data in the feature space and unable to meet the requirements for in-depth modeling of the complex interaction relationships between different modalities. Based on this, cross-modal attention analysis improves the accuracy and efficiency of feature fusion through modal interaction and feature importance scoring. The subsequently emerged generative adversarial networks (GANs) utilize discriminators to distinguish between real signals and generated features, and adversarial learning achieves the supplementation and fusion of modal information [3]. The Transformer model effectively completes tasks such as intra-modal deep encoding, cross-modal knowledge transfer, dynamic modal adaptation, and implicit missingness modeling under the condition of modal missingness [4].

2.2. Four optimization models

2.2.1 Multimodal multi-loss fusion network

The multimodal multi-loss (MMML) fusion network model aims to explore the optimal selection of cross-modal feature encoders and fusion methods, combined with multi-loss training and context modeling, to achieve a breakthrough in sentiment detection performance [5].

Its fused network module consists of a cross-attention encoder, a self-attention encoder, and a point-wise feedforward network. The cross-attention encoder achieves cross-modal projection by capturing the inter-modal dependencies, for instance, by querying the audio features (key K_{m2} and value V_{m2}) using the text features (Q_{m1}). is the vector dimension scaling factor, which prevents the dot product from being too large and thus avoiding unstable gradients; the softmax function converts the original output of the neural network into a probability distribution.

$$Attension(Q_{m1}, K_{m2}, V_{m2}) = softmax(\frac{Q_{m1}K_{m2}^{T}}{\sqrt{d_k}})V_{m2} \quad (1)$$

The self-attention encoder models the temporal correlations within a single modality, captures the dynamics within the modality, and enhances the significance of the temporal relationship of modality information. The point-wise feedforward network further optimizes the feature representation through fully connected layers and ReLU activation functions. It works in conjunction with the self-attention layer to enhance the fusion effect.

The multi-loss training adds fully connected layers at the end of each feature network to output a single-modal result. Combined with the output of the fusion network, the total loss is the sum of the losses of the three. This enhances the performance of the single-modal sub-network and simultaneously optimizes the cross-modal integration ability of the fusion network. y_m represents the modal prediction output, $target_m$ is the supervised target value, and the weight loss coefficient α_m is set to 1.

$$Loss = \sum_{m\in(a,t,f)} \alpha_m \cdot loss_{f_n}(y_m, target_m) \tag{2}$$

Furthermore, compared to the traditional context modeling approach, the method of independently processing the context and integrating it with the current utterance can support a longer context window.

Table 2. Performance Comparison between Multimodal Multi-loss Model and Typical Models

Model	Date Set	ACC_{2Has0} (%)	$F1_{Has0}$ (%)
UniMSE	CMU-MOSI	85.85	85.83
MMML	CMU-MOSI	85.91	85.85
MMML+Context	CMU-MOSI	87.51	87.45
UniMSE	CMU-MOSEI	85.86	85.79
MMML	CMU-MOSEI	86.32	86.23
MMML+Context	CMU-MOSEI	87.24	87.18

As shown in Table 2, the MMML model outperforms the current best model (UniMSE) on both the CMU-MOSI and CMU-MOSEI datasets. Furthermore, through ablation experiments on the context processing model, it was found that its performance on MMML was roughly improved by 1%~2%.

2.2.2 Multimodal prompt learning

The multimodal prompt learning model (MPLMM) is designed to efficiently generate missing modal features, optimize information fusion within and between modalities, and significantly reduce the number of training parameters [6]. This model uses three types of prompts to address the issue of modal missing data.

Firstly, the generative prompt utilizes a Conv1D convolution block and the generative prompt to generate the features of the missing modality based on the existing modal features. The missing audio feature is denoted $\hat{x}^a$ the 1D convolution block is represented as $f_{t\to\hat{a}}$, P_{Ga} is the audio generation prompt, and the concatenation operation is denoted by $f[\cdot]$.

$$(x)^a = f_{vt\to a}([P_{Ga}, f_{v\to a}(x^v), f_{t\to a}(x^t)]) \tag{3}$$

$$(x)^a = f_{t\to a}([P_{Ga}, f_{t\to a}(x^t)]) \tag{4}$$

Missing Signal Prompt marks the source of the features (real/created) to prevent incorrect features from misleading the model. P^a_{MS} represents marking the missing audio features, while P^v_{NMS} indicates that the video features are not missing.

$$(x)^a = (x)^a + P^a_{MS} \tag{5}$$

$$(x)^{v} = (x)^{v} + P_{NMS}^{v} \tag{6}$$

Missing Type Prompt (PMT), considering the correlation between modalities, sharing the processing of multi-modal missing situations by the prompts. The number of PMT prompts increases linearly with the number of modalities, effectively addressing the drawbacks of the traditional exponential growth. M_a represents the matrix of audio missing states, and P_{MT} represents the learnable original prompt matrix.

$$M_P = M_a \cdot P_{MS}^{a} + M_v \cdot P_{NMS}^{v} + M_t \cdot P_{MS}^{t} \tag{7}$$

$$P_{MT}^{'} = P_{MT} \cdot M_P \tag{8}$$

Table 3. Performance Comparison Table of MPLMM Model against Typical Models

Data Set	Method	ACC (%)	F1 (%)
CMU-MOSI	LB	64.21	67.76
	MMIN	70.84	71.22
	MPMM	69.48	70.11
	MPLMM	72.14	72.57
IEMOCAP	LB	57.74	57.42
	MMIN	65.47	-
	MPMM	65.77	65.37
	MPLMM	67.42	67.22
CH-SIMS	LB	70.11	76.24
	MMIN	71.41	76.89
	MPMM	71.26	76.85
	MPLMM	72.07	77.42

As shown in Table 3, the MPLMM model significantly outperforms the LB method, and among the other typical models, it shows the most significant improvement on the CMU-MOSI dataset, approximately by 2.5%. Furthermore, it can also maintain a lead of 1 to 2 percent on both IEMOCAP and CH-SIMS.

2.2.3 Robust multimodal missing signal framework

The Robust Multimodal Missing Signal Framework (RMSF) is designed to address the "uncertain signal missing" issue in multimodal sentiment analysis, with the aim of enhancing the model's robustness in both missing and complete modalities [7].

Firstly, the complete samples (text, audio, and video information) will be processed through the modal discard strategy (MDS) to generate 6 heterogeneous missing modal samples, simulating all possible combinations of missing data.

The RMSF model pioneered the collaborative processing of signals by three types of modules. The Hierarchical Cross-Modal (HCM) module, through "coarse-grained + fine-grained" hierarchical interaction, establishes dynamic correlations between modalities. Coarse-grained interaction utilizes all modal information to enhance the target modality, while fine-grained interaction focuses on the detailed correlations between each pair of modalities.

The adaptive feature refinement module first performs a fully connected layer projection on the joint representation output by HCM to obtain the importance scores. Subsequently, an importance matrix is generated through *softmax*, and the refined features are obtained by weighting. Finally, the Transformer encoder is used for further fusion, resulting in the final refined representation. Its purpose is to enhance the features strongly related to emotions in the available modalities and suppress noise.

The knowledge integration self-distillation module first performs knowledge similarity calculation (KSC), calculating the knowledge similarity between the complete sample and the six missing samples using the norm, and obtaining the normalized similarity through *softmax*. The second step

involves dynamic knowledge integration, where the prediction probability of the six missing samples is weighted and fused based on similarity to obtain the integrated probability. Finally, a two-way knowledge transfer is carried out. A soft objective is constructed, and the KL divergence is used to constrain the distribution consistency between the complete samples and the integrated knowledge. The aim is to utilize the complementary knowledge of heterogeneous missing samples to reconstruct the missing semantics and improve the model's adaptability to various missing scenarios.

Table 4. Performance Comparison of RMSF Model against Typical Models under the MOSI Dataset

Model	F1 in only context (%)	Avg.F1(%)	F1 in complete modality (%)
TATE	73.55	70.26	83.04
RMSF	83.75	77.86	84.79

As shown in Table 4, compared with TATE, the RMSF model increases the F1 parameter by 10.2% in the single-modal (text) scenario, but the improvement is not significant in the full-modal state, being only 1.75%. Overall, RMSF shows a very good improvement in the average F1 score (7.6%).

2.2.4 Multimodal Sentiment Analysis Based on Modal Translation

The multimodal translation and sentiment analysis model (MTMSA) fully leverages the inherent advantages of the text modality in sentiment analysis. By translating visual and audio data into text modality, it enhances the information utilization rate of low-quality modalities and fills in the missing modalities [8,9]. This model is mainly composed of the modality translation module, the Transformer module, and the common space projection.

In the preprocessing stage, the visual and audio features are projected to the same dimension as the text features through fully connected layers, facilitating cross-modal interaction. After being unified into a single dimension, the three modalities then enter the Transformer encoding stage. Through the self-attention layer and the feedforward network, they capture the temporal dependencies within the modalities and enhance the feature representation through nonlinear transformations. The Transformer decoder is the key to text translation. It uses multi-head attention and residual normalization to convert modal encoding features into text encoding features, ensuring the consistency of feature distribution. Subsequently, the JS divergence is utilized to make the translation features approximate the real text features.

The public space projection solves the problem of low efficiency in fusion caused by stitching. It uses a learning weight matrix to calculate the self-associative public space features for the translated visual, audio and original text, and then re-fuses them to lay the foundation for subsequent interaction.

After completing the feature fusion, the process will further proceed to the Transformer encoder-decoder. Through deep feature compression and feature reconstruction verification, the long-range dependencies of the fused features and the enhancement of modal semantic consistency will be addressed.

The MTMSA model adopts pre-training supervision. It provides "complete modal feature templates" through the pre-trained model, guiding the missing modal features to approach this template, without the need to distinguish specific types of missingness [10].

Table 5. Performance Comparison Table of the MTMSA Model with Typical Models

Data Set	Model	Single missing mode (Avg.ACC %)	Single missing mode (Avg.M-F1 %)	Multiple missing modalities (Avg. ACC %)	Multiple missing modalities (Avg.M-F1 %)
CMU-MOSI	MTMSA	82.64	56.35	79.94	55.11
CMU-MOSI	TATE	80.85	55.65	78.89	54.21
IEMOCAP	MTMSA	84.10	79.30	83.81	78.86
IEMOCAP	TATE	83.71	78.31	82.80	77.66

As shown in Table 5, for the CMU-MOSI dataset, the MTMSA model shows the most significant improvement in the accuracy of sentiment analysis for single-modal absence, approximately 2%. Furthermore, the MTMSA model shows an approximate 1% improvement in other performance parameters when applied to the CMU-MOSI and IEMOCAP datasets.

3. Discussion

3.1. Technical breakthrough

3.1.1 Advantages of MMML model and technological innovations

By designing independent losses for each modality sub-network, we not only optimize the overall performance but also improve the accuracy of individual modality sub-networks. The deep fusion network employs a combination of cross-attention, self-attention, and feedforward networks to replace the traditional simple concatenation or early fusion, thereby capturing fine-grained cross-modal interactions. Context modeling independently processes the refusion of context and the current statement, making more effective use of temporal information than traditional concatenated context modeling.

3.1.2 Advantages of MPLMM model and technological innovations

The generative prompt dynamically generates the missing modal features, replacing the traditional zero-value filling or retrieval-based filling methods, and more accurately captures the modal correlations. The absence of signals enables the resolution of the "confusion between true and false features" problem in traditional generation methods through modality-specific cues. The missing type hinting module generates cross-modal shared hints through projection matrices, resulting in a linear relationship between the number of hints and the number of modalities, and significantly improving parameter efficiency. It addresses the defect of traditional generated hints growing exponentially with the number of modalities.

3.1.3 Advantages of RMSF model and technological innovations

Bidirectional knowledge transfer replaces the traditional one-way distillation and enhances the reconstruction of missing semantics. Moreover, the hierarchical cross-modal interaction combined with the fine-grained innovative model digs deeper into the modal correlations compared to the traditional pairwise interaction.

3.1.4 Advantages of MTMSA model and technological innovations

Taking advantage of the inherent advantages of text in sentiment analysis, using text as the sole target modality provides a way for subsequent high-quality modality conversion. The pre-trained model supervises the generation of joint features that are close to complete modality, replacing the traditional method of designing dedicated modules for each missing scenario.

3.2. The shortcomings of the optimization model

Firstly, the generation modules of these technologies overly rely on the preprocessed features, and the quality of their features directly affects the subsequent effect of feature fusion. Secondly, convolutional networks, the transformer framework, and multi-layer granularity interaction will all reduce real-time performance. The modal translation model is overly dependent on a single modality, and its ability to reconstruct the semantic of target modality features is insufficient in extreme environments. Moreover, the framework does not have an encryption or de-sensitization module, which poses a risk of sensitive data leakage.

4. Summary

Compared with the typical technologies in the past, the optimization techniques of these four models have achieved significant performance improvements. Their approaches in feature extraction, modal fusion, and model framework are worthy of being studied and referenced in future research.

At the level of feature extraction optimization, the use of pre-trained models ensures the extraction of high-quality features. Generative prompt learning guides the dynamic reconstruction of missing modal features through trainable generative prompts.

At the modal fusion level, the two major optimization directions are hierarchical cross-modal interaction and knowledge fusion driving. The former proposes coarse and fine granularity fusion as well as missing perception cues, and simultaneously constructs inter-modal correlations and intra-modal dynamics, effectively mitigating the interference of erroneous features. The latter optimizes the traditional distillation technology and proposes joint space projection, mapping multi-modal features to a shared space for unified processing of different missing combinations, integrating the knowledge of heterogeneous missing samples, and improving generalization by re-aligning features with complete modal features through JS divergence constraints.

At the model framework level, lightweight architectures, such as the prompt learning model, only train the prompt words and the lightweight convolutional layers, significantly reducing the number of parameters. Context-aware modeling and pre-trained model supervision of these dynamic adaptation mechanisms effectively improve the accuracy of multimodal sentiment analysis.

Regarding the shortcomings of the current modal absence technology, this article makes the following outlook. Firstly, for the processing of the original signal, the current methods rely on high-quality feature extraction. It is possible to explore methods for directly processing the original signals. Secondly, multimodal alignment is crucial in the feature extraction and fusion process. When there is a serious lack, semantic alignment between modalities becomes difficult, and long-term dependency modeling can be combined with graph neural networks. Finally, future multimodal sentiment analysis will be applied in a convenient and lightweight deployment manner. Therefore, prompt learning and distillation are the key paths for deployment on edge devices.

References

[1] Diraco G, Rescio G, Siciliano P, et al. Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing. Sensors, 2023, 23(11): 5281.

[2] Wu R, Wang H, Chen H T, et al. Deep multimodal learning with missing modality: A survey. arXiv preprint arXiv:2409.07825, 2024.

[3] Ahmad Z, Jaffri Z A, Chen M, et al. Understanding GANs: Fundamentals, variants, training challenges, applications, and open problems. Multimedia Tools and Applications, 2025, 84(12): 10347-10423.

[4] Xu P, Zhu X, Clifton D A. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113-12132.

[5] Wu Z, Gong Z, Koo J, et al. Multimodal multi-loss fusion network for sentiment analysis. arXiv preprint arXiv:2308.00264, 2023.

[6] Guo Z, Jin T, Zhao Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. arXiv preprint arXiv:2407.05374, 2024.

[7] Li M, Yang D, Zhang L. Towards robust multimodal sentiment analysis under uncertain signal missing. IEEE Signal Processing Letters, 2023, 30: 1497-1501.

[8] Liu Z, Zhou B, Chu D, et al. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 2024, 101: 101973.

[9] Yang H, Zhao Y, Wu Y, et al. Large language models meet text-centric multimodal sentiment analysis: A survey. arXiv preprint arXiv:2406.08068, 2024.

[10] Zong Y, Mac Aodha O, Hospedales T. Self-supervised multimodal learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.