Mitigating Inconsistencies in Multimodal Sentiment Analysis under Uncertain Missing Modalities

Jiandian Zeng[1], **Jiantao Zhou**[1]* and **Tianyi Liu**[2]
[1] State Key Laboratory of Internet of Things for Smart City
[1] Department of Computer and Information Science, University of Macau
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University
{yb87470, jtzhou}@um.edu.mo, liutianyi@sjtu.edu.cn

Abstract

For the missing modality problem in Multimodal Sentiment Analysis (MSA), the inconsistency phenomenon occurs when the sentiment changes due to the absence of a modality. The absent modality that determines the overall semantic can be considered as a *key missing modality*. However, previous works all ignored the inconsistency phenomenon, simply discarding missing modalities or solely generating associated features from available modalities. The neglect of the key missing modality case may lead to incorrect semantic results. To tackle the issue, we propose an Ensemble-based Missing Modality Reconstruction (EMMR) network to detect and recover semantic features of the key missing modality. Specifically, we first learn joint representations with remaining modalities via a backbone encoder-decoder network. Then, based on the recovered features, we check the semantic consistency to determine whether the absent modality is crucial to the overall sentiment polarity. Once the inconsistency problem due to the key missing modality exists, we integrate several encoder-decoder approaches for better decision making. Extensive experiments and analyses are conducted on CMU-MOSI and IEMOCAP datasets, validating the superiority of the proposed method.

1 Introduction

Sentiment analysis has witnessed significant progress in the past years (Zhang et al., 2016), where the traditional textual sentiment classification has developed into more complex Multimodal Sentiment Analysis (MSA) models. Taking the phase "*Yeah, I think so.*" for instance, it is hard to read the emotion without enough lexical information, and the acoustic modality may help in the emotion recognition if available. Thus, it is crucial to combine different modalities together for accurate sentiment analysis.

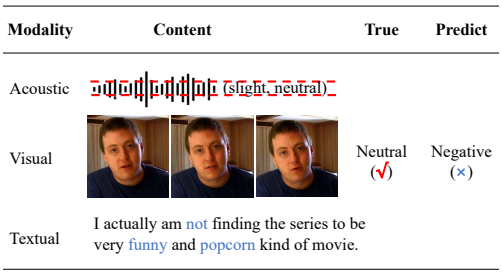


Figure 1: Case of missing the key modality, where the missing modality is marked with dotted red lines, and the semantic words are marked in blue.

So far, MSA has been well studied under the assumption that all modalities are always available. However, in reality, such a strong assumption does not always hold, and we often encounter scenarios that partial modalities could be missing. To address the missing data problem, a consequent effort has been made on recovering absent modalities. Tran et al. (2017) first identified the missing modality problem in multimodal data. More recently, several works (Suo et al., 2019; Ma et al., 2021; Zhao et al., 2021; Yuan et al., 2021; Zeng et al., 2022) focused on the missing modalities problem in an uncertain manner.

However, all of the above works ignored a vital insight that the sentiment may change when a modality is absent, resulting in the inaccurate prediction results. For instance, as shown in Fig. 1, the acoustic modality is described with the emotional tone for intuitive expression; the visual modality consists of several facial images; and the textual modality refers to the corresponding transcript. Due to the slight tone in the acoustic modality and the minor ripples in the facial features, the original emotion is neutral with full modalities. Nevertheless, once the acoustic modality is missing, the remaining sentiment is guided by the textual modality and tends to be negative. The semantics are inconsistent with or without the acoustic modality, and the absent modality can be considered as

* Corresponding Author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2924 - 2934
December 7-11, 2022 ©2022 Association for Computational Linguistics

a key missing modality. Thus, the neglect of key missing modality may lead to incorrect predictions. It is nontrivial to mark and recover the key missing modality for accurate emotion recognition in MSA. Furthermore, with the recovered features, it is still very challenging to trade off different modalities when they express different emotions.

In this paper, we tackle the above challenges by providing an ensemble solution that can accurately detect and recover features of the key missing modality. More specifically, we propose an **E**nsemble-based **M**issing **M**odality **R**econstruction (EMMR) network to handle the inconsistency problem and to further boost the performance. The proposed EMMR consists of a backbone network that utilizes an encoder-decoder structure to recover the absent modality features. Besides, to discriminate the key missing modality, we compare semantic of the recovered full modalities with the original available modalities to check their consistency. Then for mitigating the inconsistency, we aggregate Auto-Encoder (AE)-based and Transformer-based encoder-decoder approaches in an ensemble manner. Such a strategy naturally extends the feature search space, and is thus better suited to make coherent decisions. As expected and will be verified by experiments, the proposed EMMR significantly outperforms several state-of-the-art baselines on two benchmark datasets. Our major contributions are summarized as follows:

- We propose EMMR to address the inconsistency problem of missing key modality, so as to boost the performance in MSA. The code is publicly available[1].
- We integrate the AE-based and Transformer-based encoder-decoder methods for decision making to mitigate the inconsistency with better predictive performance.
- Our EMMR achieves much better performance in comparison with several state-of-the-art methods over a variety of challenging MSA datasets including CMU-MOSI and IEMOCAP.

2 Related Works

2.1 Missing Modality Problem in MSA

Regarding feature imputation strategies in MSA, previous works can be generally grouped into two categories: 1) **generative methods** (Tran et al., 2017; Vincent et al., 2008; Shang et al., 2017; Zhang et al., 2020), and 2) **joint learning methods** (Pham et al., 2019; Yuan et al., 2021).

Generative methods aim to generate new data that match the observed distributions. Variational Auto-Encoder (VAE) was proposed in (Kingma and Welling, 2014) to map the input variable to a multivariate latent distribution. Relying on GAN (Goodfellow et al., 2014), Cai et al. (2018) transformed the missing modality problem into a conditional image generation task, aiming at generating missing modality images conditioned on the existing modality. **Joint learning methods** try to learn latent representations from the observed ones. To improve the robustness of the joint representation learning, the cycle consistency strategy was applied in (Zhao et al., 2021). Also, Zeng et al. (2022) reconstructed the features of uncertain missing modalities with attached tags.

We would like to point out that the above works may make incorrect prediction without considering the inconsistency when handing the case of missing key modality. As will be clear soon, we give a comprehensive analysis in terms of inconsistency phenomenon in MSA.

2.2 Ensemble Learning

Ensemble learning (Lee et al., 2021) aims to obtain better predictive performance than a single one by combining several base models. In recent years, the ensemble technique has been applied in many NLP tasks (Li et al., 2021; Duan et al., 2021). The main idea is that it would be better to weigh and aggregate several opinions than to choose the opinion of one single individual (Sagi and Rokach, 2018). To be specific, Li et al. (2021) generated multiple candidate results with random seeds, and then trained a fusion classifier to improve the emotion recognition performance. In addition, Duan et al. (2021) developed an ensemble language model for data diversity with the technique of weight modulation. Along this line, in this paper, we aggregate several reconstruction approaches for ensemble learning to trade off different modalities when they express different emotions, and to further mitigate the inconsistency with better predictive performance.

3 Methodology

In this section, we first present the problem definition with associated notations, and then give the

[1] https://github.com/JaydenZeng/EMMR

details of all core components.

3.1 Preliminaries

Given a set of multimodal data with three modalities: $S = [X_v, X_a, X_t]$, where X_v, X_a and X_t denote visual, acoustic and textual modalities respectively. Assuming only one modality is absent, without loss of generality, we use X'_m to represent the missing modality, where $m \in \{v, a, t\}$. Formally, our problem is defined as follows: for the given triple (X_v, X_a, X_t), one modality is randomly missing. The primary task is to classify the overall sentiment (*positive*, *neutral*, or *negative*) based on the available modalities.

3.2 Backbone Network

Fig. 2 shows the backbone network based on the encoder-decoder structure. Taking the triple (X_v, X'_a, X_t) with the absent acoustic modality as an input, it is first encoded by the Multi-Head Attention (MHA) module (Vaswani et al., 2017), and then goes through two branches: 1) one is encoded by a pre-trained network which is trained with all full modalities, and 2) another goes through an encoder-decoder network to obtain the corresponding outputs, where the encoder outputs are utilized for the sentiment classification. At last, the forward similarity loss and the backward reconstruction loss are calculated to supervise the learning process of joint features.

3.3 Feature Extraction

Before being processed by the MHA module, we extract features for each modality as follows:

Visual Representations: Following (Yu et al., 2010; Zeng et al., 2022), we also adopt OpenFace2.0 toolkit (Baltrusaitis et al., 2018) to obtain 709-dimensional visual representations except data that are irrelevant attributes about the frame number, the face_id, and the timestamp, etc.

Textual Representations: For each textual utterance, the pre-trained Bert (Devlin et al., 2019) (12-layer, 768-hidden, 12-heads) is utilized to acquire 768-dimensional word vectors.

Acoustic Representations: Librosa (McFee et al., 2015) is adopted to extract 33-dimensional acoustic features, including attributes of the zero crossing rate, the Mel-Frequency Cepstral Coefficients (MFCC) and the Constant-Q Transform (CQT).

Then, all extracted modality features are encoded by the MHA module:

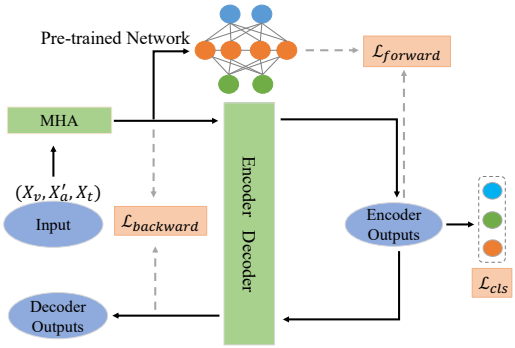


Figure 2: Structure of the backbone network.

$$E_m = \textbf{MHA}(K_m, K_m, K_m), \quad K_m \in \{X_v, X_a, X_t\}. \tag{1}$$

Afterwards, all modalities are concatenated as a whole input sequence $\mathcal{X}$:

$$\mathcal{X} = [E_v||E_a||E_t], \tag{2}$$

where $||$ is the vertically concatenating operation.

3.4 Pre-trained Network

The pre-trained network with full modalities is utilized to guide the learning process for missing modalities. To be specific, we first concatenate three full modalities, then feed them into a softmax classifier for training:

$$E_{pre} = [E_v||E_a||E_t], \quad P_{pre} = softmax(\textbf{FC}(E_{pre})). \tag{3}$$

Noting that once the model with full modalities is well trained, we fix the pre-trained network during the whole training stage.

3.5 Encoder-Decoder Network

The encoder-decoder network contains an encoder (ϕ) mapping the input ($\mathcal{X}$), and a decoder (ψ) mapping the reconstructed input ($\mathcal{X}'$), which can be defined as follows:

$$\mathcal{X} \xrightarrow{\phi} \mathcal{F}, \quad \mathcal{F} \xrightarrow{\psi} \mathcal{X}', \tag{4}$$

where $\mathcal{F}$ is the output of the encoder.

Since ensemble learning incorporates the informative knowledge from multiple models and achieves better predictive performance in an adaptive manner, it can effectively mitigate the inconsistency phenomenon. In our scheme, the AutoEncoder (AE) (Baldi, 2012), the Missing Modality

Imagination Network (MMIN) (Zhao et al., 2021), and the Transformer-based encoder-decoder model (TF) are chosen for decision making. We now introduce them one by one.

3.5.1 AE

AE is the network trained to copy its input to its output. In details, we adopt Fully Connected (FC) layers with the size of [300, 256, 128, 64, 128, 256, 300] (Please refer to the Appendix for details).

$$h_i = \begin{cases} \mathcal{X}, & i = 0 \\ ReLU(\mathbf{FC}(h_{i-1})), & 0 < i \leq 7 \end{cases}, \quad (5)$$

where the encoder output $E^{AE} = h_4$, and the decoder output $D^{AE} = h_7$.

3.5.2 MMIN

MMIN adopts the Cascade Residual Autoencoder (CRA) (Tran et al., 2017) structure with a set of Residual Autoencoders (RA). Specifically, we adopt 5 RA with the same layer settings in AE. Then the encoder output and the decoder output of the CRA can be obtained as follows:

$$\begin{aligned} D^{MMIN} &= \mathcal{X} + \sum_{i=1}^{5} \mathcal{X}_i', \\ E^{MMIN} &= \mathbf{FC}([\mathcal{F}_1||\mathcal{F}_2||...||\mathcal{F}_5]), \end{aligned} \quad (6)$$

where $\mathcal{F}_i'$ and $\mathcal{X}_i'$ are the i-th RA's encoder outputs and decoder outputs respectively.

3.5.3 TF

The Transformer architecture follows an encoder-decoder structure, which can process sequential input data effectively. With the Multi-Head Attention (MHA) mechanism and Feed-Forward Networks (FFN), the encoder output (E^{TF}) and the decoder output (D^{TF}) can be accessed:

$$\begin{aligned} E^{TF} &= \mathbf{FFN}(\mathbf{MHA}(\mathcal{X}, \mathcal{X}, \mathcal{X})), \\ D^{TF} &= \mathbf{FFN}(\mathbf{MHA}(\mathcal{F}, \mathcal{F}, \mathcal{F})), \\ \mathbf{FFN}(x) &= ReLU(W_1 x + b_1)W_2 + b_2), \end{aligned} \quad (7)$$

where W_1 and W_2 are two weight matrices, b_1 and b_2 are two learnable biases.

3.6 Ensemble

For the reconstruction of the input, we replace the missing modality with the corresponding representations in the decoder output. For instance, given the input (X_v, X_a', X_t) and the reconstructed output $(D_v^{TF}, D_a^{TF}, D_t^{TF})$, we can obtain the recovered input (X_v, D_a^{TF}, X_t). To simplify the subsequent mathematical expression, we denote the recovered input as (I_v, I_a, I_t). The sentiment of the recovered input can be acquired:

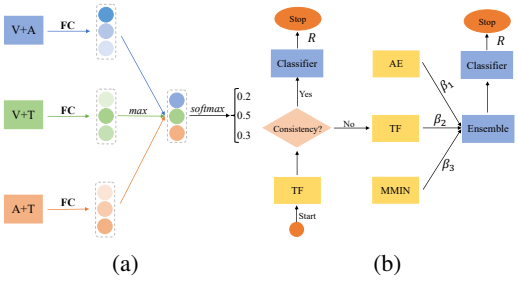


Figure 3: Illustration of ensemble methods. (a) Calculation of fusion weights for the aggregated vectors; and (b) Workflow of ensemble methods for the key missing modality.

$$L_{vat} = \mathbf{FC}(\mathbf{MHA}([I_v||I_a||I_t])). \quad (8)$$

As aforementioned, the inconsistency phenomenon occurs when the sentiment changes due to the absence of a modality in MSA. Based on this phenomenon, we utilize the inconsistency to determine whether the absent modality is crucial to the overall sentiment polarity. Specifically, we first combine every two modalities to acquire the corresponding sentiment label:

$$\begin{aligned} L_{mn} &= \mathbf{FC}(\mathbf{MHA}([I_m||I_n])), \\ m, n &\in \{v, a, t\}, m \neq n. \end{aligned} \quad (9)$$

When the sentiment label of the recovered full modalities is unequal to semantic of the remaining available modalities, the absent modality can be considered as the key missing modality. That is, in the case of (X_v, X_a', X_t), the acoustic modality is the key missing modality if $L_{vat} \neq L_{vt}$. To obtain the coherent prediction results, the inconsistency phenomenon should be mitigated. A straightforward way to handle the problem of missing key modality is voting. However, the importance of each modality is different. As shown in Fig. 3(a), we propose to assign weights according to their maximum logical values (L_k'):

$$\begin{aligned} \alpha &= softmax([L_{va}'||L_{vt}'||L_{at}']), \\ L_k' &= \max_{L_k}(softmax(L_k)), k \in \{va, vt, at\}. \end{aligned} \quad (10)$$

Then, the aggregated representation with the key missing modality can be accessed:

$$E_{key} = [\alpha_{va} L_{va} || \alpha_{vt} L_{vt} || \alpha_{at} L_{at}], \quad (11)$$

where α_{va}, α_{vt} and α_{at} are the corresponding weights calculated by Eqs. (10).

As presented in Fig. 3(b), we first feed the input into the backbone network with TF encoder-decoder. Based on the recovered features, we then check the semantic consistency between the recovered full modalities and the original available modalities. Once they are not consistent with or without the absent modality, we integrate TF, AE, and MMIN for further decision making. With the idea that the overall performance of multiple approaches in ensemble learning would be better than that of a single one, we combine three extracted features according to the corresponding attention weights. Let H be a matrix consisting of three vectors $[E_{key}^{TF} || E_{key}^{AE} || E_{key}^{MMIN}]$ produced by Eq. (11). The final representation r is formed by a weighted sum of these output vectors:

$$\begin{aligned} M &= tanh(H), \\ \beta &= softmax(w^t \cdot M), \\ r &= H \cdot \beta^T, \end{aligned} \quad (12)$$

where w is a trainable parameter vector, and T is the transpose operator. Thus, the i-th output (R_i) of our ensemble method can be formulated as following:

$$R_i = \begin{cases} r_i, & L_{vat} \neq L_{\{v,a,t\}-\{k\}} \\ \mathcal{F}_i, & otherwise \end{cases}, \quad (13)$$

where k is the absent modality, and $k \in \{v, a, t\}$.

3.7 Training Objective

The overall training objective ($\mathcal{L}_{total}$) is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{forward} + \lambda_2 \mathcal{L}_{backward}, \quad (14)$$

where $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{forward}$ is the forward differential loss, $\mathcal{L}_{backward}$ is the backward reconstruction loss, and λ_1 and λ_2 are the corresponding weights. We now introduce these loss terms in details.

Forward Differential Loss ($\mathcal{L}_{forward}$): The forward loss is calculated by the difference between the pre-trained output (E_{pre}) and the encoder output ($\mathcal{F}$), and the Kullback Leibler divergence loss function (D_{KL}) is used:

$$\mathcal{L}_{forward} = \frac{1}{2}(D_{KL}(\mathcal{F}, E_{pre}) + D_{KL}(E_{pre}, \mathcal{F})). \quad (15)$$

Backward Reconstruction Loss ($\mathcal{L}_{backward}$): For the backward loss, we aim to supervise the joint common vector reconstruction, which is calculated by the decoder output ($\mathcal{X}'$) and the processed input ($\mathcal{X}$).

$$\mathcal{L}_{backward} = \frac{1}{2}(D_{KL}(\mathcal{X}', \mathcal{X}) + D_{KL}(\mathcal{X}, \mathcal{X}')). \quad (16)$$

Classification Loss ($\mathcal{L}_{cls}$): We feed the final output R into a fully connected network with the softmax activation function for the final sentiment classification:

$$\begin{aligned} \hat{p}(y|R) &= softmax(\mathbf{FC}(R)), \\ \hat{y} &= \arg\max_y(\hat{p}(y|R)), \end{aligned} \quad (17)$$

where $\hat{y}$ is the predicted label. To be specific, we employ the standard cross-entropy loss for this classification task:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N} y_n log\hat{y}_n, \quad (18)$$

where N is the number of samples, and y_n is the true label of the n-th sample .

4 Experiments

In this section, we mainly present the experimental setup, datasets, baselines, empirical studies and observations.

4.1 Experimental Setup

Datasets: We evaluate our model on two benchmark datasets: CMU-MOSI (Zadeh et al., 2016) and IEMOCAP (Busso et al., 2008). The CMU-MOSI dataset contains 2199 segments with the sentiment score in [-3, 3]; and the IEMOCAP dataset contains 5 sessions with 151 videos. In our experiments, we report three-class (negative: [-3,0), neutral:[0], positive: (0,3]) results on CMU-MOSI, and two-class (negative:[frustration, angry, sad, fear, disappointing], positive:[happy, excited]) on IEMOCAP.

Baselines: We choose the following baselines for comparison: AE (Baldi, 2012), CRA (Tran et al., 2017) and MMIN (Zhao et al., 2021) for AE-based methods; MCTN (Pham et al., 2019),

	Models	0		0.1		0.2		0.3		0.4		0.5	
		M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
D1	AE$^\sharp$	$56.42_{\pm 0.85}$	$79.55_{\pm 0.73}$	$54.02_{\pm 0.89}$	$79.01_{\pm 0.89}$	$53.31_{\pm 0.77}$	$78.01_{\pm 0.98}$	$51.16_{\pm 1.01}$	$72.44_{\pm 0.80}$	$50.69_{\pm 0.76}$	$73.21_{\pm 1.13}$	$44.80_{\pm 0.79}$	$68.22_{\pm 1.11}$
	CRA$^\sharp$	$56.77_{\pm 0.67}$	$79.67_{\pm 0.61}$	$54.18_{\pm 0.83}$	$79.13_{\pm 0.96}$	$53.45_{\pm 0.81}$	$78.12_{\pm 0.92}$	$51.55_{\pm 0.70}$	$72.56_{\pm 0.99}$	$50.81_{\pm 0.87}$	$73.56_{\pm 1.10}$	$45.21_{\pm 0.86}$	$68.90_{\pm 1.29}$
	MCTN†	$57.17_{\pm 0.71}$	$79.48_{\pm 0.88}$	$55.23_{\pm 1.01}$	$79.52_{\pm 1.28}$	$53.80_{\pm 0.71}$	$77.38_{\pm 0.92}$	$52.17_{\pm 0.70}$	$72.47_{\pm 1.10}$	$51.39_{\pm 0.89}$	$73.69_{\pm 1.08}$	$45.27_{\pm 1.48}$	$67.75_{\pm 1.33}$
	TransM$^\sharp$	$57.79_{\pm 1.32}$	$80.14_{\pm 1.57}$	$57.34_{\pm 0.89}$	$79.55_{\pm 0.79}$	$55.09_{\pm 0.92}$	$78.30_{\pm 0.81}$	$52.55_{\pm 1.18}$	$72.89_{\pm 0.95}$	$52.33_{\pm 0.79}$	$72.12_{\pm 0.90}$	$45.43_{\pm 1.39}$	$68.04_{\pm 1.87}$
	MMIN†	$60.39_{\pm 0.78}$	$82.20_{\pm 0.63}$	$57.69_{\pm 1.02}$	$81.78_{\pm 0.98}$	$55.33_{\pm 0.77}$	$80.15_{\pm 0.95}$	$53.47_{\pm 0.81}$	$79.17_{\pm 1.13}$	$52.32_{\pm 0.98}$	$76.28_{\pm 1.33}$	$48.87_{\pm 1.41}$	$70.55_{\pm 1.79}$
	TATE†	$58.27_{\pm 0.52}$	$84.88_{\pm 0.78}$	$58.21_{\pm 0.69}$	$84.32_{\pm 0.57}$	$55.29_{\pm 1.21}$	$81.25_{\pm 0.95}$	$55.08_{\pm 0.79}$	$80.56_{\pm 1.12}$	$54.01_{\pm 0.86}$	$79.95_{\pm 1.39}$	$51.55_{\pm 1.48}$	$73.82_{\pm 1.44}$
	Ours	$\mathbf{68.08}_{\pm 0.78}$	$\mathbf{85.93}_{\pm 0.65}$	$\mathbf{67.17}_{\pm 0.72}$	$\mathbf{85.24}_{\pm 0.83}$	$\mathbf{66.41}_{\pm 1.15}$	$\mathbf{84.37}_{\pm 0.76}$	$\mathbf{64.21}_{\pm 1.04}$	$\mathbf{82.81}_{\pm 0.94}$	$\mathbf{62.55}_{\pm 1.38}$	$\mathbf{81.77}_{\pm 1.26}$	$\mathbf{60.75}_{\pm 1.78}$	$\mathbf{78.81}_{\pm 1.65}$
D2	AE$^\sharp$	$76.23_{\pm 0.51}$	$82.07_{\pm 0.71}$	$75.22_{\pm 0.68}$	$80.15_{\pm 0.46}$	$75.17_{\pm 0.60}$	$77.60_{\pm 0.97}$	$73.88_{\pm 0.57}$	$77.21_{\pm 0.73}$	$77.10_{\pm 0.81}$	$75.85_{\pm 1.08}$	$67.19_{\pm 0.77}$	$76.29_{\pm 0.99}$
	CRA$^\sharp$	$77.10_{\pm 0.66}$	$82.11_{\pm 0.78}$	$75.93_{\pm 0.84}$	$80.68_{\pm 0.53}$	$75.22_{\pm 0.44}$	$77.73_{\pm 1.04}$	$74.55_{\pm 0.60}$	$78.19_{\pm 0.71}$	$79.55_{\pm 0.66}$	$76.08_{\pm 0.93}$	$67.66_{\pm 0.64}$	$76.44_{\pm 1.28}$
	MCTN†	$78.55_{\pm 0.48}$	$82.12_{\pm 0.72}$	$77.69_{\pm 0.56}$	$80.79_{\pm 0.66}$	$75.21_{\pm 0.50}$	$78.22_{\pm 0.94}$	$74.50_{\pm 0.87}$	$78.48_{\pm 0.70}$	$71.72_{\pm 0.48}$	$76.25_{\pm 1.11}$	$68.05_{\pm 0.77}$	$76.54_{\pm 1.25}$
	TransM$^\sharp$	$79.55_{\pm 0.66}$	$82.57_{\pm 0.71}$	$77.49_{\pm 0.92}$	$80.72_{\pm 0.74}$	$76.28_{\pm 0.55}$	$80.29_{\pm 0.68}$	$75.79_{\pm 0.59}$	$78.45_{\pm 0.62}$	$71.77_{\pm 0.84}$	$77.13_{\pm 0.88}$	$68.32_{\pm 1.28}$	$76.59_{\pm 1.36}$
	MMIN†	$80.79_{\pm 0.78}$	$83.41_{\pm 0.83}$	$78.82_{\pm 0.69}$	$82.49_{\pm 0.94}$	$76.90_{\pm 0.86}$	$81.15_{\pm 0.72}$	$76.55_{\pm 0.64}$	$80.40_{\pm 1.08}$	$73.11_{\pm 1.32}$	$78.38_{\pm 0.89}$	$70.51_{\pm 0.76}$	$77.41_{\pm 1.22}$
	TATE†	$81.22_{\pm 0.76}$	$85.29_{\pm 0.77}$	$80.05_{\pm 0.63}$	$85.18_{\pm 0.71}$	$79.19_{\pm 0.96}$	$84.05_{\pm 0.83}$	$78.43_{\pm 0.70}$	$83.18_{\pm 0.79}$	$76.71_{\pm 1.11}$	$82.69_{\pm 0.97}$	$74.39_{\pm 1.26}$	$81.99_{\pm 1.55}$
	Ours	$\mathbf{83.58}_{\pm 0.47}$	$\mathbf{86.51}_{\pm 0.55}$	$\mathbf{82.47}_{\pm 0.63}$	$\mathbf{85.37}_{\pm 0.58}$	$\mathbf{79.55}_{\pm 0.61}$	$\mathbf{84.92}_{\pm 0.73}$	$\mathbf{79.18}_{\pm 0.47}$	$\mathbf{83.88}_{\pm 0.63}$	$\mathbf{78.01}_{\pm 0.92}$	$\mathbf{82.68}_{\pm 1.03}$	$\mathbf{76.76}_{\pm 1.12}$	$\mathbf{82.02}_{\pm 1.27}$

Table 1: Performance of all baselines, where D1 and D2 denote the CMU-MOSI and the IEMOCAP datasets respectively. The best results are in bold. The results with $^\sharp$ are reproduced, and the results with † are re-generated under the same settings.

and TransM (Wang et al., 2020) for translation-based methods; TATE (Zeng et al., 2022) and the proposed EMMR for transformer-based methods. $Accuracy$ (ACC) and $Macro - F1$ (M-F1) are used to measure the performance of the models.

The detailed implementation, dataset statistics, and hyper-parameter settings are available in the attached Appendix.

4.2 Overall Results

Table 1 shows the qualitative results with all baselines. Our proposed EMMR achieves the best results on all settings, especially about 8.54% to 11.12% improvement in terms of M-F1 on the CMU-MOSI dataset. The present results are significant due to the fact that three ensemble approaches can well handle the inconsistency problem when missing a key modality, so as to further improve the robustness. Besides, the performance has a gradual drop with more absent samples when the missing ratio increases from 0 to 0.5. We also find that MCTN and TransM achieve better performance than AE and CRA, implying that cyclic translations can better fuse the multimodal information from multiple modalities. In addition, TATE and EMMR outperform other baselines due to the strong learning ability of the transformer structure. Another observation is that our proposed EMMR still performs well when nearly half of samples are missing, which is caused by the reason that three ensemble methods can combine their predictions in a complementary manner.

4.3 Effects of Different Settings

In this subsection, we first conduct the ablation studies to better understand the influence of different modules. Afterwards, we further evaluate the performance of our model by replacing several core components with alternatives.

1) Ablation study: We evaluate our model with several settings: a) using only one modality; b) using two modalities; c) removing the pre-trained network; and d) removing the backward reconstruction module.

According to the results given in Table 2, it can be seen that the performance drops sharply with a single modality, especially when removing the textual modality. However, similar reductions are not observed when the visual modality is missing. These results suggest that the textual modality may dominate the overall sentiment. Besides, one striking result to emerge from the data is that the performance improves when combing two modalities, indicating that multiple modalities can boost the performance by learning complementary features from each other. In addition, referring to the last two lines, the performance decreases about 9.97% to 14.39% with respect to M-F1 and about 7.60% to 9.12% on ACC when the pre-trained network is removed, showing the importance of the forward guidance. Meanwhile, further analysis suggests that the backward reconstruction module also provides a good supervision for the final joint representation learning.

2) Effects of different ensemble methods: We now examine the effectiveness of different ensemble methods. For the comparison purpose, we conduct experiments with several settings: a) using only the backbone network, b) combing two ensemble methods, c) combing three ensemble methods with the maximum operation, and d) combing three ensemble methods with the average operation.

As can be seen in Table 3, although the back-

Modules	0		0.1		0.2		0.3		0.4		0.5	
	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC
V	$40.54_{\pm 0.86}$	$56.85_{\pm 0.91}$	-	-	-	-	-	-	-	-	-	-
A	$41.23_{\pm 0.75}$	$60.88_{\pm 1.21}$	-	-	-	-	-	-	-	-	-	-
T	$57.32_{\pm 0.59}$	$77.48_{\pm 0.61}$	-	-	-	-	-	-	-	-	-	-
V+A	$42.32_{\pm 0.57}$	$61.39_{\pm 0.66}$	$41.28_{\pm 0.43}$	$60.27_{\pm 0.87}$	$39.78_{\pm 0.67}$	$59.37_{\pm 0.81}$	$39.47_{\pm 0.79}$	$59.63_{\pm 0.55}$	$38.48_{\pm 0.95}$	$58.66_{\pm 0.70}$	$38.01_{\pm 1.45}$	$57.39_{\pm 1.55}$
V+T	$59.67_{\pm 0.55}$	$81.45_{\pm 0.62}$	$58.85_{\pm 0.72}$	$80.49_{\pm 0.63}$	$57.63_{\pm 0.79}$	$79.56_{\pm 0.85}$	$56.14_{\pm 0.97}$	$78.82_{\pm 0.69}$	$55.86_{\pm 0.78}$	$77.67_{\pm 1.21}$	$53.27_{\pm 1.43}$	$76.99_{\pm 1.75}$
A+T	$59.95_{\pm 0.61}$	$81.89_{\pm 0.52}$	$59.12_{\pm 0.66}$	$80.87_{\pm 0.49}$	$58.55_{\pm 0.68}$	$80.11_{\pm 0.59}$	$57.42_{\pm 0.83}$	$79.41_{\pm 0.60}$	$56.78_{\pm 0.87}$	$78.21_{\pm 0.85}$	$55.39_{\pm 0.71}$	$77.43_{\pm 1.49}$
V+A+T	$\mathbf{68.08}_{\pm 0.78}$	$\mathbf{85.93}_{\pm 0.65}$	$\mathbf{67.17}_{\pm 0.72}$	$\mathbf{85.24}_{\pm 0.83}$	$\mathbf{66.41}_{\pm 1.15}$	$\mathbf{84.37}_{\pm 0.76}$	$\mathbf{64.21}_{\pm 1.04}$	$\mathbf{82.81}_{\pm 0.94}$	$\mathbf{62.55}_{\pm 1.38}$	$\mathbf{81.77}_{\pm 1.26}$	$\mathbf{60.75}_{\pm 1.78}$	$\mathbf{78.81}_{\pm 1.65}$
-w/o $\mathcal{L}_{forward}$	$55.11_{\pm 0.64}$	$77.83_{\pm 0.76}$	$53.78_{\pm 0.56}$	$76.12_{\pm 0.97}$	$52.27_{\pm 0.87}$	$75.38_{\pm 0.76}$	$51.83_{\pm 0.73}$	$74.54_{\pm 1.28}$	$51.41_{\pm 0.85}$	$72.77_{\pm 1.01}$	$50.78_{\pm 1.21}$	$71.21_{\pm 1.38}$
-w/o $\mathcal{L}_{backward}$	$57.47_{\pm 0.34}$	$79.56_{\pm 0.48}$	$56.12_{\pm 0.41}$	$78.17_{\pm 0.65}$	$54.79_{\pm .63}$	$77.28_{\pm 0.49}$	$53.27_{\pm 0.55}$	$76.13_{\pm 0.68}$	$52.19_{\pm 0.88}$	$75.43_{\pm 0.96}$	$51.96_{\pm 1.43}$	$73.29_{\pm 1.77}$

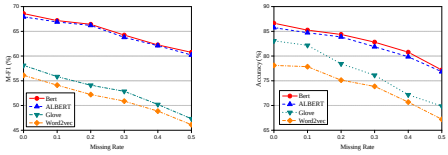
Table 2: Comparison of different modules on CMU-MOSI.

Settings	0		0.2		0.4	
	M-F1	ACC	M-F1	ACC	M-F1	ACC
TF	$58.43_{\pm 0.72}$	$82.55_{\pm 0.66}$	$55.28_{\pm 0.85}$	$80.64_{\pm 0.91}$	$52.79_{\pm 1.13}$	$77.21_{\pm 1.25}$
TF+AE	$60.71_{\pm 0.59}$	$82.91_{\pm 0.54}$	$57.24_{\pm 0.76}$	$80.92_{\pm 0.88}$	$55.75_{\pm 1.08}$	$78.10_{\pm 1.10}$
TF+MMIN	$62.44_{\pm 0.61}$	$83.29_{\pm 0.85}$	$59.85_{\pm 0.70}$	$81.78_{\pm 0.88}$	$56.49_{\pm 0.97}$	$78.99_{\pm 1.26}$
Max	$65.98_{\pm 0.71}$	$83.85_{\pm 0.62}$	$63.87_{\pm 0.88}$	$82.01_{\pm 0.80}$	$59.57_{\pm 1.12}$	$79.89_{\pm 1.45}$
Average	$66.83_{\pm 0.65}$	$84.17_{\pm 0.51}$	$64.19_{\pm 0.73}$	$82.96_{\pm 0.80}$	$60.84_{\pm 1.25}$	$80.60_{\pm 1.16}$
Ours	$\mathbf{68.08}_{\pm 0.78}$	$\mathbf{85.93}_{\pm 0.65}$	$\mathbf{66.41}_{\pm 1.15}$	$\mathbf{84.37}_{\pm 0.76}$	$\mathbf{62.55}_{\pm 1.38}$	$\mathbf{81.77}_{\pm 1.26}$

Table 3: Results of different ensemble methods.

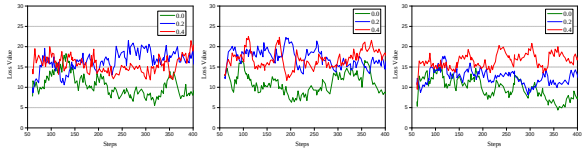
Ratio	2-class		4-class		7-class	
	M-F1	ACC	M-F1	ACC	M-F1	ACC
0	$\mathbf{83.58}_{\pm 0.47}$	$\mathbf{86.51}_{\pm 0.55}$	$\mathbf{56.88}_{\pm 0.58}$	$\mathbf{62.12}_{\pm 0.72}$	$\mathbf{38.55}_{\pm 0.61}$	$\mathbf{48.29}_{\pm 0.70}$
0.1	$82.47_{\pm 0.63}$	$85.37_{\pm 0.58}$	$55.08_{\pm 0.63}$	$58.05_{\pm 0.71}$	$36.77_{\pm 0.82}$	$47.39_{\pm 0.95}$
0.2	$79.55_{\pm 0.66}$	$84.92_{\pm 0.73}$	$53.85_{\pm 0.87}$	$57.49_{\pm 0.70}$	$36.21_{\pm 0.77}$	$45.01_{\pm 0.85}$
0.3	$79.18_{\pm 0.47}$	$83.88_{\pm 0.63}$	$51.05_{\pm 0.72}$	$56.83_{\pm 0.69}$	$35.91_{\pm 0.66}$	$44.38_{\pm 0.80}$
0.4	$78.01_{\pm 0.92}$	$82.68_{\pm 1.03}$	$48.52_{\pm 0.62}$	$55.89_{\pm 0.81}$	$34.83_{\pm 1.02}$	$43.26_{\pm 1.31}$
0.5	$76.76_{\pm 1.12}$	$82.02_{\pm 1.27}$	$46.99_{\pm 0.86}$	$55.23_{\pm 1.27}$	$33.82_{\pm 1.20}$	$43.10_{\pm 1.46}$

Table 4: Results of multiple classes on IEMOCAP.



(a) M-F1 (b) ACC

Figure 4: Results of different word embeddings. (a) M-F1; and (b) ACC.



(a) MAE (b) Cosine (c) Ours

Figure 5: Training loss curves of different settings. (a) MAE; (b) Cosine; and (c) Ours.

bone network with TF achieves competitive performance, there is still improvement when combining AE and MMIN. The reason may be that ensemble learning combines knowledge from multiple models to achieve better predictive performance. Besides, TF+MMIN outperforms than TF+AE, implying that MMIN can extract better modality features than AE. Compared to the average operation, our weighted fusion method improves about 1.71% to 2.25% with respect to M-F1 and about 1.17% to 1.76% on ACC, validating the effectiveness of the weighted fusion mechanism.

3) Effects of different word embeddings: As aforementioned, the textual modality may dominate the overall sentiment, and we now evaluate the performance of different word embedding models. To this end, we choose Word2vec (Mikolov et al., 2013), Glove (Pennington et al., 2014) and ALBERT (Lan et al., 2020) as alternative methods to the pre-trained Bert, and evaluate the respective prediction performance. Here, we set the embedding size as 128 in Word2vec and choose the cased 840B tokens of 300 dimension in Glove. All settings share the same parameters for a fair comparison. As presented in Fig. 4, different embedding models have significant effect on the overall performance, where Bert-based methods achieve better results while the Word2vec model is the worst. These results altogether provide an important insight that Bert embeddings result in better word semantic correlations, as it is trained from a large amount of text corpus.

4) Effects of multiple classes: We would also like to observe the performance of multiple classes on IEMOCAP. Apart from the general 2-class results, the happy, angry, sad and neutral emotions are chosen as the 4-class experiment, and the extra frustration, excited, and surprise emotions are selected as the 7-class experiment. Table 4 reveals that there has been a sharp drop in both M-F1 and ACC with more emotion categories. More specifically, the performance of the 7-classes experiment drops by almost half due to the confusion of multiple categories, and the model is hard to classify them correctly. Further efforts are needed to boost the performance under scenarios of multiple classes.

5) Effects of different losses: We further ex-

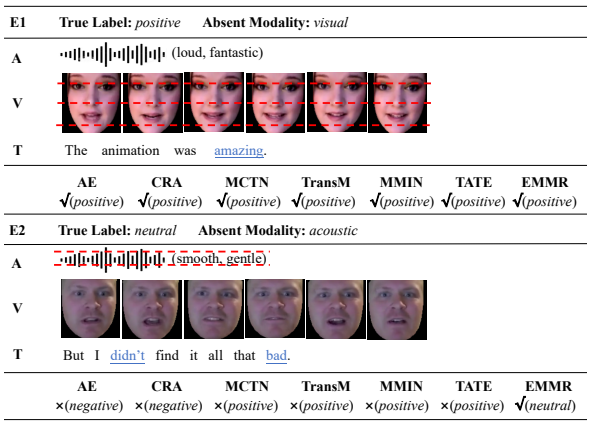


Figure 6: Two cases of the test data, along with their predicted categories by all baselines, where × (or ✓) means that the predicted category is wrong (or correct).

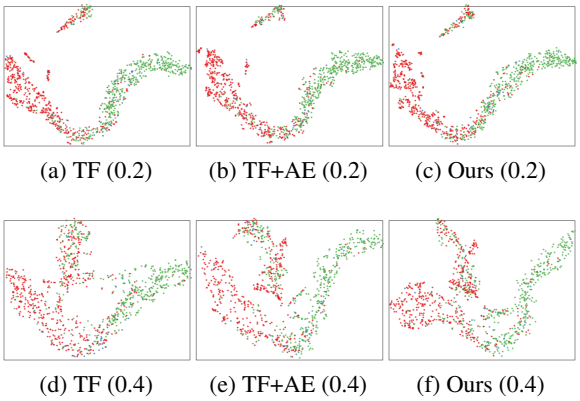


Figure 7: Visualization of different ensemble methods. The top (a)-(c) are with 20% missing rate; and the bottom (d)-(e) are with 40% missing rate.

plore the effects of different losses. For the comparison purpose, we choose the MAE loss and the cosine loss as alternative methods to the KL loss. Fig. 5 presents the training loss curves (steps ranging from 50 to 300) on the CMU-MOSI dataset, including three missing rates of 0, 0.2, and 0.4. It can be observed that the training loss curves in our method (Fig. 5(c)) fluctuate relatively smoother than other two loss settings (Fig. 5(a)-(b)), showing the good convergence of our setting. Besides, the training loss curves become more fluctuating with the increment of the missing rate, especially when the missing rate is 0.4. Compared to the cosine similarity loss and the MAE loss, our KL divergence loss leads to the smaller minimum loss values of 4.89. We then conclude that the KL divergence loss provides a good assessment of the similarity between two probability distributions.

4.4 Case Study

To better understand in which conditions the proposed method works, we present several challenging cases for further analyses. To this end, two examples are given in Fig. 6, where blue words with underline potentially express sentiment polarity, and the missing modality is marked with dotted red lines.

From the figure, we can find: 1) In **E1**, all models generate correct results though the visual modality is missing. Due to the strong guidance of the textual word *"amazing"*, the positive polarity is obviously expressed. This case reveals that the conventional approaches can be well-performed when existing modalities express the same explicit semantics. 2) In **E2**, the textual modality expresses positive polarity, while the visual modality tends to be negative because of the frown and close lips on facial features. It is really hard to determine the polarity when the acoustic modality is missing. Specifically, AE and CRA misclassify the emotion as negative, and the other approaches except EMMR all predict positive sentiment in terms of the dominance of the textual modality. In contrast, our method (EMMR) first discriminates whether the inconsistency phenomenon exists, then integrates three methods to acquire better decisions in a complementary manner.

4.5 Visualization

To further demonstrate the learning ability of different ensemble models, we adopt the T-SNE toolkit to present the learned joint representations in Fig. 7. To be specific, we visualize about 1000 vectors with three ensemble settings on CMU-MOSI, where the red, the blue, and the green colors denote negative, neutral and positive respectively. As can be observed, in Fig. 7(a)-(c), all learned vectors are generally clustered into three categories with TF as the backbone network. Besides, there are less outliers with more ensemble approaches, due to the reason that the errors of one single model can be compensated by other models. Such phenomenon also agrees with the observations from Fig. 7(c)-(d). Furthermore, the clusters in the red and the green colors are more discrete with bigger missing rate. We then conclude that the model is hard to converge with too many absent samples and thus degrades the performance.

5 Conclusion

In this paper, we focus on mitigating the inconsistency phenomenon when a key modality is absent in MSA. The proposed EMMR first learns features from remaining modalities via a backbone encoder-decoder network. Then, we discriminate the key modality by checking the semantic consistency between the recovered full modalities and the original available modalities. Afterwards, three ensemble approaches based on the backbone encoder-decoder network are utilized to make decisions when the inconsistency phenomenon exists. Experimental results and analyses are provided to demonstrate the effectiveness of our scheme compared with several state-of-the-art methods. Future research will focus on aggregating different ensemble approaches for a comprehensive analysis.

6 Limitations

We would like to discuss the detailed limitations in this section. As aforementioned, we integrate three different encoder-decoder approaches for decision making when the inconsistency phenomenon exists. Although it is nontrivial to select the right ensemble methods and to utilize them correctly, the model for ensemble learning can be expensive in terms of both time and space. As can be seen in the attached Appendix, a comprehensive comparison of the overall parameters and the testing time has been carried out, which motivates us to further optimize the proposed model effectively.

7 Acknowledgements

This work was supported in part by Macau Science and Technology Development Fund under SKLIOTSC-2021-2023, 0072/2020/AMJ, and 0022/2022/A1; in part by Research Committee at University of Macau under MYRG2018-00029-FST, MYRG2019-00023-FST, MYRG2020-00101-FST and MYRG2022-00152-FST; in part by Natural Science Foundation of China under 61971476; and in part by Alibaba Group through Alibaba Innovative Research Program.

References

Pierre Baldi. 2012. Autoencoders, unsupervised learning, and deep architectures. In *Proc. ICML Workshop Unsupervised and Transf. Learn.*, pages 37–49.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Int. Conf. Auto. Face Amst. Recognit.*, pages 59–66. IEEE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, 42(4):335–359.

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pages 1158–1166.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. North Am. Ch. Assoc. Comput. Linguist. Hun. Lang. Tech.*, pages 4171–4186.

Zhibin Duan, Hao Zhang, Chaojie Wang, Zhengjue Wang, Bo Chen, and Mingyuan Zhou. 2021. Enslm: Ensemble language model for data diversity by semantic clustering. In *Proc. Assoc. Comput. Linguist. Int. J. Conf. Nat. Lang. Process.*, volume 1, pages 2954–2967.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Adv. Neural Inf. Process. Syst.*, volume 27, pages 1–9.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proc. Int. Conf. Learn. Represent.*, pages 1–14.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *Proc. Int. Conf. Learn. Represent.*, pages 1–17.

Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2021. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *Proc. Int. Conf. on Mach. Learn.*, pages 6131–6141.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021. Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In *Proc. Empir. Methods Nat. Lang. Process.*, pages 3935–3941.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proc. AAAI Conf. Artif. Intell.*, volume 35, pages 2302–2310.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in

python. In *Proc. Python Sci. Conf.*, volume 8, pages 18–25.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. Workshop Int. Conf. Learn. Representations*, pages 1–12.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. Empir. Methods Nat. Lang. Process.*, pages 1532–1543.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proc. AAAI Conf. Artif. Intell.*, volume 33, pages 6892–6899.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8(4):e1249.

Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. 2017. Vigan: Missing view imputation with generative adversarial networks. In *IEEE Int. Conf. Big Data*, pages 766–775.

Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. 2019. Metric learning on healthcare data with incomplete modalities. In *Proc. Int. J. Conf. Artifi. Intell.*, pages 3534–3540.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1405–1414.

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 *Adv. Condens. Matter Phys.*, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proc. Int. Conf. Mach. Learn.*, pages 1096–1103.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proc. Web Conf.*, pages 2514–2520.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2010. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proc. Assoc. Comput. Linguist.*, pages 3718–3727.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proc. ACM Int. Conf. Multimedia*, pages 4400–4407.

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

Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Ret.*, pages 1–10.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. 2020. Deep partial multi-view learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2402–2415.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proc. AAAI Conf. Artif. Intell.*, pages 3087–3093.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proc. Assoc. Comput. Linguist. Int. Jt. Conf. Nat. Lang. Process.*, pages 2608–2618.

A Network Structure

We present the network structure of the pre-trained model with full modalities in Fig. 8(a), the ensemble AE-based encoder-decoder network in Fig. 8(b), and the Missing Modality Imagination Network (MMIN) in Fig. 8(c).

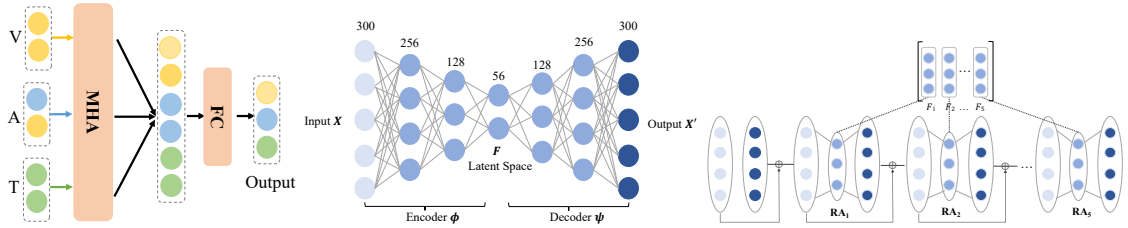
To be specific, in Fig 8(a), three modalities are first encoded by the Multi-Head Attention (MHA) module, and then are concatenated for classification. In Fig. 8(b), the hidden sizes of full connected layers are in [300, 256, 128, 64, 128, 256, 300]. In Fig. 8(c), we adopt 5 Residual Autoencoders (RA) with the same layer settings in AE, where the encoder outputs are obtained by concatenating the latent space of 5 RA blocks.

B Implementation Details

All experiments are carried out on a Linux server (Ubuntu 18.04.1) with a Intel(R) Xeon(R) Gold 5120 CPU, 128G RAM, 8 Nvidia 2080TI and 2 Nvidia 3090 GPUs.

B.1 Datasets Distributions

The detailed distributions on CMU-MOSI and IEMOCAP are shown in Table 5. Besides, the distributions of multiple classes on IEMOCAP are presented in Table 6.



(a) Pre_trained Network (b) AE (c) MMIN

Figure 8: Network Structure. (a) Pre_trained network; (b) AE; and (c) MMIN.

Dataset		Pos.	Neu.	Neg.	Total
CMU-MOSI	Train	833	81	866	1780
	Val	92	8	100	200
	Test	98	7	94	199
IEMOCAP	Train	1006	-	2510	3516
	Val	301	-	827	1128
	Test	329	-	848	1177

Table 5: Detailed distributions on two datasets.

Dataset		hap.	ang.	sad	neu.	fru.	exc.	sur.	Total
4-Classes	Train	349	659	653	1042	-	-	-	2703
	Val	122	225	208	321	-	-	-	876
	Test	124	219	223	345	-	-	-	911
7-Classes	Train	345	702	638	1024	1107	607	66	4489
	Val	111	205	218	345	368	214	22	1483
	Test	139	196	228	339	374	220	19	1515

Table 6: Detailed distributions on multiple classes on IEMOCAP.

B.2 Hyper-parameters

Following a standardized procedure, we tune our model by the grid-searching on the training set. Adam is adopted to minimize the total loss. The batch size is 32, the loss weight is set to 0.1, and these parameters are summarized in Table 7.

C Memory and Running Time

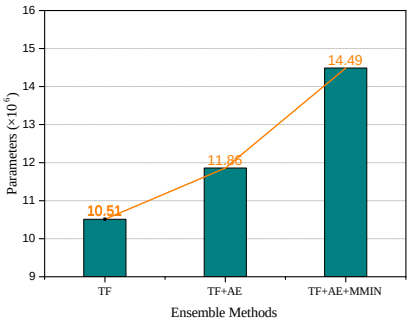


Figure 9: Parameters of different ensemble approaches.

For the memory utilization, Fig. 9 presents the parameters of different ensemble approaches. As can be observed, the number of parameters dramatically increase when integrating MMIN. The reason is that MMIN contains 5 residual auto-encoders, which are memory costly.

Description	Symbol	Value
Batch size	b	32
Epoch number	e	20
Dropout rate	p	0.3
Hidden size	d	300
Missing rate	η	[0, 0.5]
Learning rate	lr	0.001
Maximum textual length	n_t	25
Maximum visual length	n_v	100
Maximum acoustic length	n_a	150
Loss weights	λ_1, λ_2	0.1

Table 7: Detailed parameter settings in our experiments.

Dataset		2080Ti	3090
Training	TF	1826.44	1156.62
	TF+AE	1885.51	1192.76
	TF+AE+MMIN	1975.18	1246.20
Testing	TF	57.88	24.10
	TF+AE	62.39	27.17
	TF+AE+MMIN	78.08	30.41

Table 8: Running time (s) of different ensemble approaches.

As for the training and testing time, we show the detailed statistics in Table 8. Specifically, we report the training time at 10 epochs and the testing time for the test dataset on 2080Ti and 3090 GPUs respectively. It can be seen that the testing time is acceptable though the training time varies considerably during training. Besides, compared to the 2080Ti GPU, the 3090 GPU spends less time due to its stronger computational capability. Although the proposed EMMR boosts the performance, it can be expensive regarding to both time and space, motivating us to further optimize the model effectively.