Missing Modality meets Meta Sampling (M^3S): An Efficient Universal Approach for Multimodal Sentiment Analysis with Missing Modality

Haozhe Chi Minghua Yang Junhao Zhu Guanhong Wang Gaoang Wang*
Zhejiang University-University of Illinois at Urbana-Champaign Institute,
Zhejiang University, China
{haozhe.20, minghua.20, junhao.20, gaoangwang}@intl.zju.edu.cn
guanhongwang@zju.edu.cn

Abstract

Multimodal sentiment analysis (MSA) is an important way of observing mental activities with the help of data captured from multiple modalities. However, due to the recording or transmission error, some modalities may include incomplete data. Most existing works that address missing modalities usually assume a particular modality is completely missing and seldom consider a mixture of missing across multiple modalities. In this paper, we propose a simple yet effective meta-sampling approach for multimodal sentiment analysis with missing modalities, namely Missing Modality-based Meta Sampling (M^3S). To be specific, M^3S formulates a missing modality sampling strategy into the modal agnostic meta-learning (MAML) framework. M^3S can be treated as an efficient add-on training component on existing models and significantly improve their performances on multimodal data with a mixture of missing modalities. We conduct experiments on IEMOCAP, SIMS and CMU-MOSI datasets, and superior performance is achieved compared with recent state-of-the-art methods.

1 Introduction

Multimodal sentiment analysis (MSA) aims to estimate human mental activities by multimodal data, such as a combination of audio, video, and text. Though much progress has been made recently, there still exist challenges, including missing modality problem. In reality, missing modality is a common problem due to the errors in data collection, storage, and transmission. To address the issue with missing modality in MSA, many approaches have been proposed (Ma et al., 2021c; Zhao et al., 2021; Ma et al., 2021b; Parthasarathy and Sundaram, 2020; Ma et al., 2021a; Tran et al., 2017).

In general, methods that address the missing modality issue usually only consider the situation where a certain input modality is severely damaged.

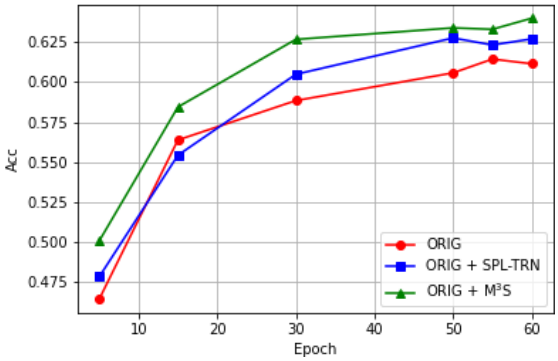


Figure 1: M^3S helps MMIN model achieve superior performance.

The strategies of these proposed methods can be divided into three categories: 1) Designing new architectures with a reconstruction network to recover missing modality with the information from other modalities (Ma et al., 2021c; Ding et al., 2014); 2) Formulating innovative and efficient loss functions to tackle missing modality (Ma et al., 2021a, 2022); 3) Improving the encoding and embedding strategies from existing models (Tran et al., 2017; Cai et al., 2018).

In the MSA tasks, most of the proposed methods focus on the situation where certain modalities are completely missing and the other modalities are complete. However, due to the transmission or collection errors, each modality may contain partial information based on a certain missing rate, while existing methods seldom consider this type of scenario and they are not suitable to be applied directly in this situation. Besides, our experiments also verify the inefficacy of existing methods in such a challenging situation, which is demonstrated in Section 5.

To address the aforementioned problem, in this paper, we propose a simple yet effective solution to the **M**issing **M**odality problem with **M**eta **S**ampling in the MSA task, namely M^3S. To be specific, M^3S combines the augmented missing modality trans-

*Corresponding author.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 121–130
November 20–23, 2022.

form in sampling, following the model-agnostic meta-learning (MAML) framework (Finn et al., 2017). M^{3}S maintains the advantage of meta-learning and makes models easily adapt to data with different missing rates. M^{3}S can be treated as an efficient add-on training component on existing models and significantly improve their performances on multimodal data with a mixture of missing modalities. We conduct experiments on IEMOCAP (Busso et al., 2008), SIMS (Yu et al., 2020) and CMU-MOSI (Zadeh et al., 2016) datasets and superior performance is achieved compared with recent state-of-the-art (SOTA) methods. A simple example is shown in Figure 1, demonstrating the effectiveness of our proposed M^{3}S compared with other methods. More details are provided in the experiment section.

The main contributions of our work are as follows:

- We formulate a simple yet effective meta-training framework to address the problem of a mixture of partial missing modalities in the MSA tasks.
- The proposed method M^{3}S can be treated as an efficient add-on training component on existing models and significantly improve their performances on dealing with missing modality.
- We conduct comprehensive experiments on widely used datasets in MSA, including IEMOCAP, SIMS, and CMU-MOSI. Superior performance is achieved compared with recent SOTA methods.

2 Related Work

2.1 Emotion Recognition

Emotion recognition aims to identify and predict emotions through these physiological and behavioral responses. Emotions are expressed in a variety of modality forms. However, early studies on emotion recognition are often single modality. Shaheen et al. (2014) and Calefato et al. (2017) present novel approaches to automatic emotion recognition from text. Burkert et al. (2015) and Deng et al. (2020) conduct researches on facial expressions and the emotions behind them. Koolagudi and Rao (2012) and Yoon et al. (2019) exploit acoustic data in different types of speeches for emotional recognition and classification tasks. Though much progress has been made for emotion recognition with single modality data, how to combine information from diverse modalities has become an interesting direction in this area.

2.2 Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA) is a popular area of research in the present since the world we live in has several modality forms. When the dataset consists of more than one modality information, traditional single modality methods are difficult to deal with. MSA mainly focuses on three modalities: text, audio, and video. It makes use of the complementarity of multimodal information to improve the accuracy of emotion recognition. However, the heterogeneity of data and signals bring significant challenges because it creates distributional modality gaps. Hazarika et al. (2020) propose a novel framework, MISA, which projects each modality to two distinct subspaces to aid the fusion process. And Hori et al. (2017) introduce a multimodal attention model that can selectively utilize features from different modalities. Since the performance of a model highly depends on the quality of multimodal fusion, Han et al. (2021b) construct a framework named MultiModal InfoMax (MMIM) to maximize the mutual information in unimodal input pairs as well as obtain information related to tasks through multimodal fusion process. Besides, Han et al. (2021a) make use of an end-to-end network Bi-Bimodal Fusion Network (BBFN) to better utilize the dynamics of independence and correlation between modalities. Due to the unified multimodal annotation, previous methods are restricted in capturing differentiated information. Yu et al. (2021) design a label generation module based on the self-supervised learning strategy. Then, joint training the multimodal and unimodal tasks to learn the consistency and difference. However, limited by the pre-processed features, the results show that the generated audio and vision labels are not significant enough.

2.3 Missing Modality Problem

Compared with unimodal learning method, multimodal learning has achieved great success. It improves the performance of emotion recognition tasks by effectively combining the information from different modalities. However, the multimodal data may have missing modalities in reality due to a variety of reasons like signal transmission error

and limited bandwidth. To deal with this problem, Ma et al. (2021b) propose an efficient approach based on maximum likelihood estimation to incorporate the knowledge in the modality-missing data. Nonetheless, the more complex scenarios like missing modalities exist in both training and testing phases are not involved. What's more, recent studies aim to capture the common information in different types of training data and leverage the relatedness among different modalities (Ma et al., 2021a; Tran et al., 2017; Parthasarathy and Sundaram, 2020; Wagner et al., 2011). To solve the problem that modalities will be missing is uncertain, Zhao et al. (2021) put forward a unified model: Missing Modality Imagination Network (MMIN). Ma et al. (2021c) utilize a new method named SMIL that leverages Bayesian meta-learning to handle the problem that modalities are partially severely missing, *e.g.*, 90% training examples may have incomplete modalities.

3 Methodology

3.1 Problem Description

The multimodal sentiment analysis aims at predicting the sentiment labels $\mathcal{Y}$ based on the model $f(\mathcal{X};\boldsymbol{\theta})$ given the multimodal data $\mathcal{X}$. We consider the input data with three modalities, *i.e.* $\mathcal{X} = (\mathcal{A}, \mathcal{V}, \mathcal{L})$, where $\mathcal{A}$, $\mathcal{V}$ and $\mathcal{L}$ represents audio, video and linguistic data, respectively. In this paper, we tackle the missing modality issue, where each modality can include missing data.

Algorithm 1 Meta-Sampling Training

Input: Multimodal dataset $(\mathcal{X} = (\mathcal{A}, \mathcal{V}, \mathcal{L}), \mathcal{Y})$; number of iterations K for inner loop; inner learning rate α; outer learning rate β; estimation model $f(\cdot\,;\boldsymbol{\theta})$; model's loss function $l\,(f, \mathcal{Y})$.

while not converged **do**
  Sample batch of data $\mathcal{X}_1$ and $\mathcal{X}_2$ from $\mathcal{X}$.
  Get $\tilde{\mathcal{X}}_1 = \mathcal{T}(\mathcal{X}_1\,;\mathcal{F})$ and $\tilde{\mathcal{X}}_2 = \mathcal{T}(\mathcal{X}_2\,;\mathcal{F})$.
  Set $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$
  Meta-train:
  for $n = 0$ to $K-1$ **do**
    $\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n - \alpha\nabla_{\theta_n}\, l\left(f(\tilde{\mathcal{X}}_1;\boldsymbol{\theta}_n), \mathcal{Y}_1\right)$
  end for
  $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_K$
  Meta-update:
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta\nabla_{\theta^*}\, l\left(f(\tilde{\mathcal{X}}_2;\boldsymbol{\theta}^*), \mathcal{Y}_2\right)$
end while

3.2 Augmented Missing Modality Transform

Given a sample $\boldsymbol{X}_i = (\boldsymbol{A}_i, \boldsymbol{V}_i, \boldsymbol{L}_i)$ from $\mathcal{X}$, we use a augmented transform $\mathcal{T}(\boldsymbol{X}_i\,;\mathcal{F})$ to generate a random sample with missing data based on a distribution $\mathcal{F}$. Specifically, for each modality $m \in \{a, v, l\}$, we define a missing ratio $r_m \in [0, 1]$, where a, v and l stands for audio, video and linguistic modality, respectively. For the encoded feature in each modality m, we replace the values between $[\lambda_m, \lambda_m + k_m - 1]$ with zeros, where k_m represents the number of missing values with $k_m = \lfloor T_m \cdot r_m \rfloor$ and T_m is the dimension of the encoded feature. λ_m is sampled from the uniform distribution, *i.e.*, $\lambda_m \sim \mathcal{U}(0, T_m - k_m)$. As a result, the augmented sample with missing modality can be obtained by $\tilde{\boldsymbol{X}}_i = \mathcal{T}(\boldsymbol{X}_i\,;\mathcal{F})$, where $\mathcal{F}$ represents the composition of uniform distributions for each individual modality.

3.3 Training with Meta-Sampling

Our M^{3}S follows MAML training framework (Finn et al., 2017) with augmentation sampling. For each training iteration, we adopt the following steps.

First, we sample two independent batch of data, $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$, based on the augmented missing modality transforms, $\mathcal{T}(\mathcal{X}_1\,;\mathcal{F})$ and $\mathcal{T}(\mathcal{X}_2\,;\mathcal{F})$, where the missing rate for each modality is determined by the sampling distribution $\mathcal{F}$. $\tilde{\mathcal{X}}_1$ and $\tilde{\mathcal{X}}_2$ are used as tasks from support set and query set, respectively, in the meta-learning.

Then, in the meta-train process, the model's parameter $\boldsymbol{\theta}$ is updated using gradient descent based on the loss function $l\left(f(\tilde{\mathcal{X}}_1;\boldsymbol{\theta}), \mathcal{Y}_1\right)$ with the inner learning rate α for each iteration n as follows:

$$\boldsymbol{\theta}_{n+1} \leftarrow \boldsymbol{\theta}_n - \alpha\nabla_{\theta_n}\, l\left(f(\tilde{\mathcal{X}}_1;\boldsymbol{\theta}_n), \mathcal{Y}_1\right), \quad (1)$$

where $\mathcal{Y}_1$ is the set of sentiment labels of $\tilde{\mathcal{X}}_1$, and the loss function $l\left(f(\tilde{\mathcal{X}}_1;\boldsymbol{\theta}), \mathcal{Y}_1\right)$ is determined by loss used in each base model (*i.e.*, MMIM, MISA, Self-MM, MMIN. See Section 4.2 for more details). The meta-train process is conducted for K iterations. We denote $\boldsymbol{\theta}_K$ as $\boldsymbol{\theta}^*$.

Finally, we use the query set $\tilde{\mathcal{X}}_2$ and its set of sentiment labels $\mathcal{Y}_2$ in the outer loop meta-update step. The model parameters are updated with the learning rate β as follows:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta\nabla_{\theta^*}\, l\left(f(\tilde{\mathcal{X}}_2;\boldsymbol{\theta}^*), \mathcal{Y}_2\right). \quad (2)$$

The whole algorithm in general case is shown

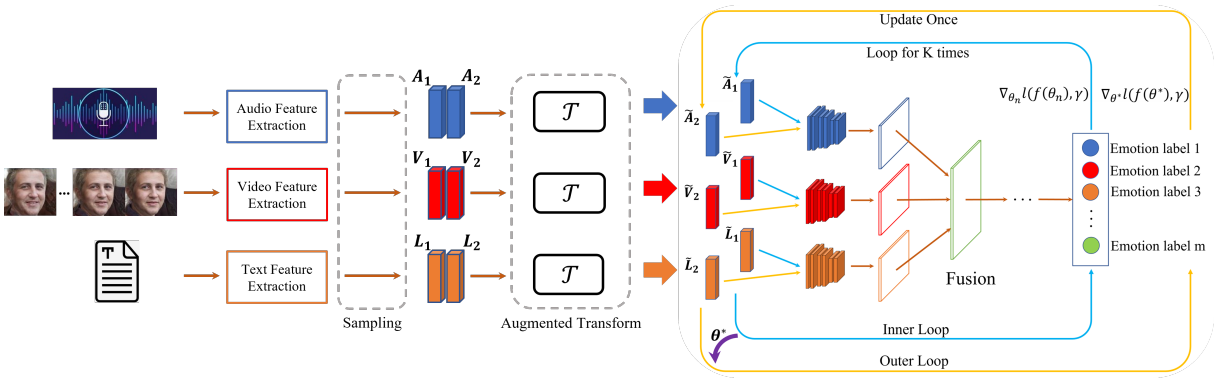


Figure 2: The Overall Architecture of M^{3}S. We first use augmented transform to generate two batches of data for features from each modality. Then the meta-train and meta-update are conducted on the two batches of data to learn the model parameters $\boldsymbol{\theta}$.

in Algorithm 1 and Figure 2 illustrates the meta-sampling training process.

4 Experiment Setup

In this section, we present the setup of our experiments, including the used datasets, baseline methods, evaluation metrics, and implementation details of the proposed method.

4.1 Datasets

We conduct our experiments on the following three datasets, *i.e.*, IEMOCAP (Busso et al., 2008), SIMS (Yu et al., 2020) and CMU-MOSI (Zadeh et al., 2016). The statistics of the datasets are reported in Table 1.

- **IEMOCAP** comprises of several recorded videos in 5 conversation sessions, and each session contains many scripted plays and dialogues. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions, which provided detailed information about their facial expressions and hand movements.

- **SIMS** dataset is a multimodal sentiment analysis benchmark containing 2281 video clips from various sources (*i.e.*, movies, shows, TV serials, etc.). SIMS contains fine-grained annotations of different modalities and includes people's natural expressions in video clips. And each sample in SIMS dataset is labeled with a score from -1 to 1, standing for sentiment response (*i.e.*, from strongly negative to strongly positive).

Dataset	Train	Valid	Test	All
SIMS	1368	456	457	2281
MOSI	1284	229	686	2199
IEMOCAP	4446	3342	3168	10956

Table 1: Statistics of the Used Datasets

- **CMU-MOSI** has 2199 video segments in total, which are sliced from 93 YouTube videos. The videos address a large array of topics like books, products, and movies. In these video segments, 89 narrators show their opinions on different topics. Most of the speakers are around 20-30 years old. They all express themselves in English, although they come from different countries.

4.2 Baseline Methods

We use four recent SOTA methods for comparison in the experiments. The methods include MMIM (Han et al., 2021b), MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021) and MMIN (Zhao et al., 2021), which are summarized as follows.

† **MMIM** helps mutual information reach maximum and maintains information related to tasks during the process of multimodal fusion, which shows significant results in multimodal sentiment analysis tasks.

† **MISA** is a novel model in emotion recognition that represents modality more effectively and improves the fusion process significantly.

† **Self-MM** has novel architecture containing several innovative modules (like a module for

Method	Self-MM (SIMS)				MMIN (IEMOCAP)		
	MAE	Corr	Acc-2	F1-Score	Acc	Uar	F1-Score
ORIG	0.5171	0.3918	0.7291	0.6980	0.6136	0.6403	0.6049
ORIG + SPL-TRN	**0.5049**	0.4080	0.7392	0.7102	0.6357	0.6518	0.6235
ORIG + M^{3}S	0.5053	**0.4091**	**0.7405**	**0.7119**	**0.6398**	**0.6536**	**0.6296**
Δ_{ORIG}	↓ 0.0118	↑ 0.0173	↑ 0.0114	↑ 0.0139	↑ 0.0262	↑ 0.0133	↑ 0.0247
Method	MISA (MOSI)			-	MMIM (MOSI)		
	MAE	Corr	Acc-7	-	MAE	Corr	Acc-7
ORIG	0.8886	0.7349	0.3863	-	0.7175	0.7883	0.4592
ORIG + SPL-TRN	**0.8279**	**0.7355**	0.4155	-	0.7126	0.7825	0.4650
ORIG + M^{3}S	0.8393	0.7346	**0.4282**	-	**0.7014**	**0.7985**	**0.4852**
Δ_{ORIG}	↓ 0.0493	↓ 0.0003	↑ 0.0419	-	↓ 0.0161	↑ 0.0102	↑ 0.0260

Table 2: Results of four baseline models with different training methods applied. Input and test data both have missing rates between 40% and 60%. ORIG stands for original model; SPL-TRN stands for sampling-training. Δ_{ORIG} presents the improved performance based on original model that M^{3}S has achieved.

label generation) and reaches brilliant results in multimodal sentiment analysis tasks.

† **MMIN** handles the problem that input data has uncertain modalities completely missing and achieves superior results under various missing modality conditions.

4.3 Evaluation Metrics

Following the four baseline methods mentioned above, we use the following evaluation metrics, including mean absolute error (MAE), Pearson correlation (Corr), binary classification accuracy (Acc-2), weighted F1 score (F1-Score), accuracy score (Acc), unweighted average recall (Uar), and seven-class classification accuracy (Acc-7). Acc-7 denotes the ratio of predictions that are in the correct interval among the seven intervals ranging from -3 to 3. For all metrics, higher values show better performance except for MAE.

4.4 Implementation Details

Hyperparameter Settings. The settings of inner learning rate, outer learning rate and batch size $\{\alpha, \beta, \text{batch\_size}\}$ are as follows: MMIN {2e-4, 1e-4, 256}; MMIM {1e-3, 1e-3, 32}; MISA {1e-4, 1e-4, 128}; For Self-MM, the learning rate for three modalities $\{\mathcal{A}, \mathcal{V}, \mathcal{L}\}$ is {5e-3, 5e-3, 5e-5}, and the batch size is 32.

Feature Extraction Details. Following the baseline methods, we adopt the extracted features as the input for each modality. The feature extraction methods on each modality $\{\mathcal{A}, \mathcal{V}, \mathcal{L}\}$ are listed as follows: MMIN {OpenSMILE-"IS13_ComParE" (Eyben et al., 2010), DenseNet (Huang et al., 2017) trained on FER+ corpus (Barsoum et al., 2016), BERT (Devlin et al., 2018)}; Self-MM, MMIM, MISA {sLSTM (Hochreiter and Schmidhuber, 1997), sLSTM, BERT}.

Experimental Details. We use Adam as the optimizer for all four baseline models. The training epoch for {MMIN, MMIM, MISA} is $\{60, 40, 500\}$. Self-MM adopts the "early stop" strategy to obtain the best result. Therefore, its training epoch is unfixed. In Section 5.1, We compare the performance of three different training methods dealing with missing modalities in our experiment results: 1) original model's training method (ORIG), where the missing rate of each sample is fixed along the training process during different epochs; 2) original model with Sampling-Training strategy applied (ORIG + SPL-TRN), which adopts augmented sampling without meta-learning process, as illustrated in Section 3.2; 3) original model with M^{3}S added on (ORIG + M^{3}S), which is the proposed method.

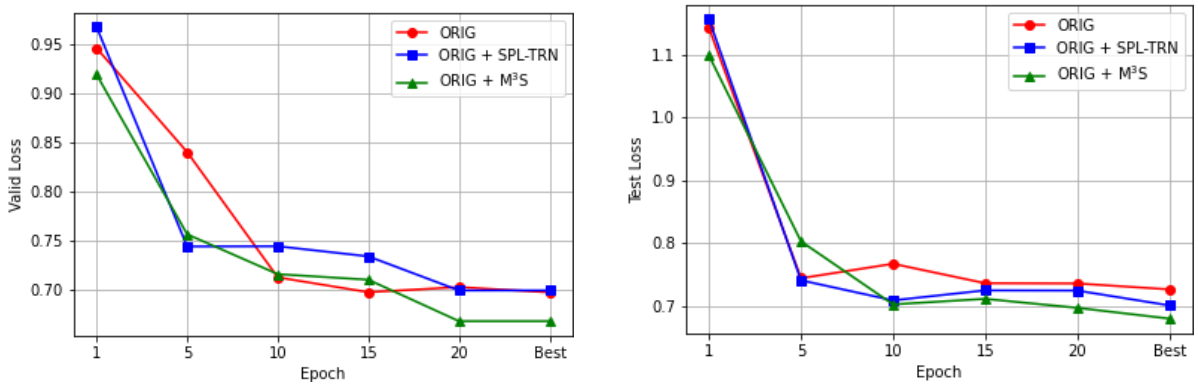
5 Results and Analysis

5.1 Main Results

Built on the baseline models, we conduct experiments with the proposed M^{3}S method and show its effectiveness in Table 2. The missing rate is set as the medium rate, between 40% and 60%. Since M^{3}S can be an add-on component to existing methods with the capability of dealing with missing

Input Missing Rate	Method	MMIN (IEMOCAP)			MMIM (MOSI)		
		Acc	Uar	F1-Score	MAE	Corr	Acc-7
60% ∼ 80%	ORIG	0.5849	0.5915	0.5748	**0.7132**	**0.7905**	0.4577
	ORIG + SPL-TRN	0.5812	0.5901	0.5689	0.7268	0.7867	0.4549
	ORIG + M^{3}S	**0.5900**	**0.6026**	**0.5764**	0.7208	0.7890	**0.4588**
	Δ_{ORIG}	↑ 0.0051	↑ 0.0111	↑ 0.0016	↑ 0.0076	↓ 0.0015	↑ 0.0011
40% ∼ 60%	ORIG	0.6136	0.6403	0.6049	0.7175	0.7883	0.4592
	ORIG + SPL-TRN	0.6357	0.6518	0.6235	0.7126	0.7825	0.4650
	ORIG + M^{3}S	**0.6398**	**0.6536**	**0.6296**	**0.7014**	**0.7985**	**0.4852**
	Δ_{ORIG}	↑ 0.0262	↑ 0.0133	↑ 0.0247	↓ 0.0161	↑ 0.0102	↑ 0.0260
20% ∼ 40%	ORIG	0.6192	0.6453	0.6078	0.7129	0.7893	0.4694
	ORIG + SPL-TRN	0.6335	**0.6513**	0.6221	0.7218	0.7832	0.4665
	ORIG + M^{3}S	**0.6367**	0.6504	**0.6266**	**0.7049**	**0.7923**	**0.4838**
	Δ_{ORIG}	↑ 0.0175	↑ 0.0051	↑ 0.0188	↓ 0.0080	↑ 0.0030	↑ 0.0144

Table 3: Results on MMIN and MMIM under three different missing rate levels. Test data have the same range of missing rates as input data.



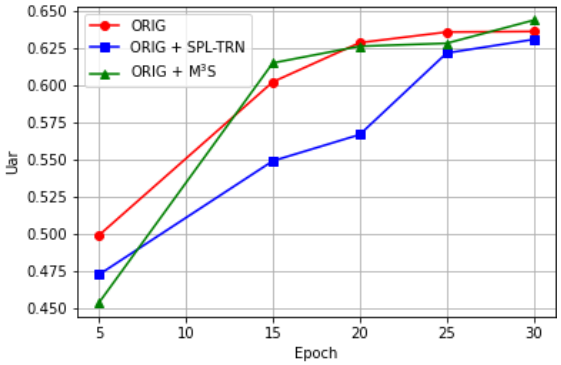
(a) Valid Loss (b) Test Loss

Figure 3: Validation and testing losses of three methods along training built on the MMIM Model.

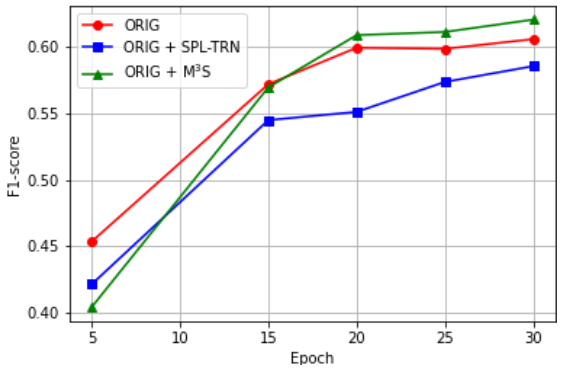
modality, we compare M^{3}S with Sampling-Training (SPL-TRN) and four original baseline methods. For all the testing datasets, M^{3}S achieves superior performance in almost all evaluation metrics compared with the original baseline methods, as expected. Since SPL-TRN only adopts augmented sampling without meta-learning process, it achieves worse performance than our M^{3}S method in most of the experiments. This result demonstrates that the meta-sampling training process can better learn the common knowledge from other modalities to deal with the missing information. It also verifies that meta-training can better utilize the information from random augmentations. As a matter of fact, with the help of M^{3}S, MMIN model achieves the highest Acc, highest Uar, and highest F1-Score. Also, built upon the other three baselines (Self-MM, MISA, MMIM), M^{3}S helps in reaching the lowest MAE, highest Corr, and highest Acc in most situations, which shows the efficiency and universality of M^{3}S.

5.2 Studies of Various Missing Rates

To verify the effectiveness of methods on different missing rates, we conduct experiments on two datasets by varying the input missing rate to three levels (*i.e.*, 20%-40%, 40%-60%, and 60%-80%). Results in Table 3 show that for nearly all the cases, our method M^{3}S outperforms ORIG and ORIG+SPL-TRN methods. Specifically, when input missing rate falls within the range 40%-60%, ORIG+M^{3}S shows the greatest increment in all metrics, which shows that M^{3}S achieves the most significant effect on models with medium missing level.



(a) Uar



(b) F1-Score

Figure 4: Uar and F1-Score of three methods along training built on the MMIN Model.

5.3 Convergence Comparison

As is shown in Figure 3(a) and 3(b), we plot the process of MMIM model's loss decline. It is clearly shown in plots that M^{3}S helps original model converge to the lowest loss after 10 to 15 epochs of training. As shown in Figure 4(a) and Figure 4(b), we also select MMIN model and plot its convergence process because the trend of its metrics changes more obviously. These two figures, along with Figure 1 show the characteristic of our method: although M^{3}S does not show strong competitiveness in the first few epochs, with the progress of training, M^{3}S helps model achieve faster growth of various metrics and finally converge to a higher result.

5.4 Adaptation across Different Missing Rates

In order to further discover the efficiency of our method in helping models adapt to different missing rates, we conduct experiments with testing rates different from input rates. As shown in Table 4, compared to ORIG method, we can see that M^{3}S significantly improves nearly all metrics by at least 1%. It is worth noticing that a large missing rate

MMIN	ORIG	ORIG + SPL-TRN	ORIG + M^{3}S	Δ_{ORIG}
Acc	0.6035	0.6152	**0.6206**	↑ 0.0171
Uar	**0.6281**	0.6166	0.6140	↓ 0.0141
F1-Score	0.5953	0.6023	**0.6072**	↑ 0.0119
MMIM	**ORIG**	**ORIG + SPL-TRN**	**ORIG + M^{3}S**	Δ_{ORIG}
MAE	0.7201	0.7412	**0.7025**	↓ 0.0176
Corr	0.7794	0.7695	**0.7884**	↑ 0.0090
Acc-7	0.4534	0.4461	**0.4825**	↑ 0.0291

Table 4: Results on MMIN (IEMOCAP) and MMIM (MOSI), where input data have missing rates 40%-60% and test data have missing rates 60%-80%.

(60%-80%) is adopted in the testing, and M^{3}S achieves much better performance than the other two methods. For example, the Acc-7 of M^{3}S on MOSI dataset is over 3.6% higher than the one of ORIG+SPL-TRN method, demonstrating the capability of M^{3}S when different modalities have large missing information.

5.5 Further Discussion and Limitations

The qualitative results and ablation study above show that M^{3}S significantly helps baseline models improve their performance on inputs with various missing rates. However, when we apply M^{3}S to Self-MM model and conduct experiments on CMU-MOSI dataset, we find that the results show little difference from the original model's result. Besides, from Table 2 we know that M^{3}S improves Self-MM's performance on SIMS dataset significantly. Hence we assume that this is because Self-MM model has good adaptability to CMU-MOSI dataset but not SIMS dataset when both datasets have a mixture of missing across modalities. Therefore, some models may show adaptivity to certain datasets. And M^{3}S may not significantly improve the model's performance on those datasets that model is already quite adaptive to.

Also, as shown in Table 3, it's revealed that when inputs have a large missing rate (60%-80%), M^{3}S becomes limited in improving evaluation metrics. We attribute this to the change of sampling range. That is, when inputs have missing rates no more than 60%, we can create sufficient augmented missing data to perform M^{3}S. However, when inputs have large missing rates, we can only get augmented data with missing rates restricted to a smaller range. Thus we get a smaller sampling range containing large missing rate data, which makes M^{3}S limited.

But in general, M^{3}S method is recommended as it

P-value of t-test	Self-MM (SIMS)				MMIN (IEMOCAP)		
	MAE	Corr	Acc-2	F1-Score	Acc	Uar	F1-Score
$P(T \leq t)$	0.1959	0.0384	0.0018	0.0615	0.0007	7.95E-5	0.0005
P-value of t-test	MISA (MOSI)			-	MMIM (MOSI)		
	MAE	Corr	Acc-7	-	MAE	Corr	Acc-7
$P(T \leq t)$	0.0473	0.1873	0.0405	-	0.0277	0.1971	0.0263

Table 5: Two-tailed significance test (t-test) of M^{3}S.

is easy to be added on different models and efficient in improving models' performance on multimodal sentiment analysis tasks most of the time, especially when input data has a medium missing rate. As shown in Table 5, nearly all evaluation metrics' P-value is smaller than 0.05 in the significance test, indicating significant improvement when M^{3}S is applied.

6 Conclusion and Future Work

In this paper, we focus on a challenging problem, *i.e.*, multimodal sentiment analysis on a mixture of missing across modalities, which was seldom studied in the past. We propose a simple yet effective method called M^{3}S to handle the problem. M^{3}S is a meta-sampling training method that follows the MAML framework and combines the sampling strategy for augmented transforms. M^{3}S maintains the advantages of meta-learning and helps SOTA models achieve superior performance on various missing input modalities.

In the experiments, we show that our method M^{3}S improves four baselines' performance and helps them adapt to inputs with various missing rates. Furthermore, M^{3}S is easy to realize in different multimodal sentiment analysis models. In future work, we plan to investigate how to better combine M^{3}S with other training methods and extend the method to other multimodal learning tasks.

Ethical Considerations

Our proposed method aims to help improve the performance of different SOTA methods on data with various missing rates. All experiments we conduct are based on the open public datasets (Section 4.1) and pretraining baseline methods (Section 4.2). When applying our method in experiments, there is minimal risk of privacy leakage. Furthermore, since our method is an add-on component for different baselines, it is safe to apply it as long as the baseline model provides adequate protection for privacy.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62106219). We would like to thank the anonymous reviewers for their valuable and detailed suggestions.

References

Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 279–283.

Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. 2015. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*.

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. *Language Resources and Evaluation*, 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 *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1158–1166.

Fabio Calefato, Filippo Lanubile, and Nicole Novielli. 2017. Emotxt: A toolkit for emotion recognition from text. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, pages 79–80. IEEE.

Didan Deng, Zhaokang Chen, and Bertram E Shi. 2020. Multitask emotion recognition with incomplete labels. In *2020 15th IEEE International Conference*

on Automatic Face and Gesture Recognition, pages 592–599. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengming Ding, Shao Ming, and Yun Fu. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

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. *arXiv preprint arXiv:2109.00412*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

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

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4193–4202.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International Journal of Speech Technology*, 15(2):99–117.

Fei Ma, Shao-Lun Huang, and Lin Zhang. 2021a. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *2021 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE.

Fei Ma, Xiangxiang Xu, Shao-Lun Huang, and Lin Zhang. 2021b. Maximum likelihood estimation for multimodal learning with missing modality. *arXiv preprint arXiv:2108.10513*.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021c. Smil: Multimodal learning with severely missing modality. *arXiv preprint arXiv:2103.05677*.

Srinivas Parthasarathy and Shiva Sundaram. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, pages 383–392. IEEE.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1405–1414.

Johannes Wagner, Elisabeth Andre, Florian Lingenfelser, and Jonghwa Kim. 2011. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2822–2826. IEEE.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *arXiv preprint arXiv:2102.04830*.

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.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. 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, pages 2608–2618.