Transformer-based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis

Ziqi Yuan[12]*, Wei Li[12]*, Hua Xu[12]†, Wenmeng Yu[12]
[1]State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Beijing National Research Center for Information Science and Technology(BNRist), Beijing 100084, China
yzq21@mails.tsinghua.edu.cn,w-li17@mails.tsinghua.edu.cn
xuhua@tsinghua.edu.cn,ywm18@tsinghua.org.cn

ABSTRACT

Improving robustness against data missing has become one of the core challenges in Multimodal Sentiment Analysis (MSA), which aims to judge speaker sentiments from the language, visual, and acoustic signals. In the current research, translation-based methods and tensor regularization methods are proposed for MSA with incomplete modality features. However, both of them fail to cope with random modality feature missing in non-aligned sequences. In this paper, a transformer-based feature reconstruction network (TFR-Net) is proposed to improve the robustness of models for the random missing in non-aligned modality sequences. First, intra-modal and inter-modal attention-based extractors are adopted to learn robust representations for each element in modality sequences. Then, a reconstruction module is proposed to generate the missing modality features. With the supervision of SmoothL1Loss between generated and complete sequences, TFR-Net is expected to learn semantic-level features corresponding to missing features. Extensive experiments on two public benchmark datasets show that our model achieves good results against data missing across various missing modality combinations and various missing degrees.

CCS CONCEPTS

• **Information systems** → *Multimedia and multimodal retrieval*; *Web log analysis*; Web applications; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

multimodal sentiment analysis, transformer, feature reconstruction, data missing

*Equal Contribution.
†Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475585

ACM Reference Format:
Author1, Author2, Author3, Author4. 2021. Transformer-based Feature Reconstruction Network for Robust Multimodal Sentiment Analysis. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3474085.3475585

1 INTRODUCTION

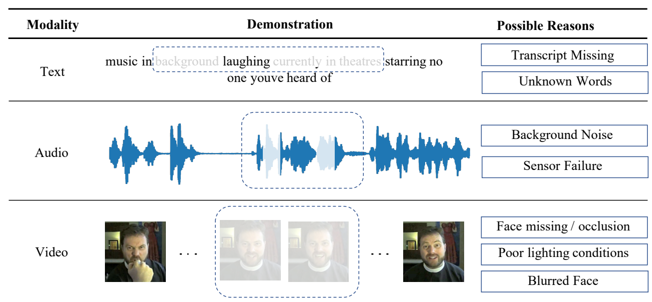

Figure 1: Factors that may lead to random modality missing problem.

With the abundance of user-generated online content, MSA has recently become an active area in Natural Language Processing (NLP) [16, 20]. Using manually aligned complete information, including transcript language, audio and vision, previous work has achieved significant improvement on MSA task. However, user-generated videos in the wild are often imperfect. First, the receptors for different modalities may have variable receiving frequency, which leads to unaligned nature. Second, as shown in Figure 1, many inevitable factors, such as corrupted noise or sensor failure in user-generated videos, may result in the failure of modality feature extractors.

Under the above circumstances, the need for a model which can deal with Random Modality Feature Missing (RMFM) arises. As a result, constructing models which can handle RMFM in MSA is still an open research. The core challenge in MSA with RMFM lies in the sparse semantics in incomplete modality sequence, leading to the difficulty in extracting robust modality representation. To the best of our knowledge, current works do not devote efforts to regenerate the missing semantics in modality sequences. Instead, they directly use the incomplete modality sequences with the missing penalty to learn joint fusion representations. However, due to the lack of semantic information in missing sequences, the improvement is limited.

Encoder-decoder framework is first proposed as a sequence to sequence method in Neural Machine Translation (NMT) [5], and soon adapt to many multimodal translate tasks such as Image/Video Caption [18, 25] and Visual Question Answering [4]. In the current research, this framework is also used in multimodal representation learning, as its ability to generate hidden representation capturing the shared semantics from both source and target sequences [9]. Motivated by it, we form an encoder-decoder framework to reproduce the semantics for the missing elements. Specifically, the encoder takes the incomplete modality sequence as input, and perform inter-modal and cross-modal attention mechanism to extract the semantics of the modality sequence. The proposed decoder tries to project the enriched sequence representation into the input space. By minimizing reconstruction loss between the generated sequence and the complete modality sequence, the model learns to extract semantics from the incomplete modality sequences. Furthermore, we utilize a late fusion strategy to fusion the enriched modality sequences and make sentiment predictions.

In brief, the contributions of our work can be summarized as follows:

- As far as we know, this article is the first work focusing on the multimodal sentiment analysis task with random missing in non-aligned modality sequences and proposes a complete and reasonable evaluation model to evaluate the robustness of multimodal incomplete data.
- This paper proposes a novel method based on the encoder-decoder framework to guide modality feature extraction regenerating sequence features with missing part of the semantics.
- The proposed model performs well in experiments on benchmark multi-modal sentiment classification datasets. From the experimental results, we can conclude that TFR-Net is a general framework that is flexible to deal with the incompleteness of non-aligned features in various modalities and various degrees.

2 RELATED WORK

In this part, we mainly introduce the related works of this paper. Firstly, the traditional unaligned MSA models and the MSA models in which the modality missing problem is considered will be explained. Next, we will briefly introduce the transformer and Bert language model, by which our proposed model can be more effective. Finally, the generative models for dealing with various kinds of incomplete data will be described.

2.1 Multimodal Sentiment Analysis

MSA aims to predict the people's sentiment from the video, audio, and text of the utterances. The models like MFN [30] and EF-LSTM [27] can work on aligned multimodal data, which means the frames of audio and vision have explicit correspondence with the words in the text modality. To deal with more practical scenarios, MSA models are gradually expanding to the area of unaligned multimodal data inputs. TFN [29] and LMF [13] use tensor-based method to get joint representation for utterances. MulT [22] utilize cross-modal transformers to handle the unaligned multimodal data. MISA [10] learns modality-invariant and -specific representation for each modality to improve the fusion process. However, in those models, the extra processing for missing multimodal data does not exist.

There are several works in MSA which aims to solve the missing data problem in MSA. MCTN [17] uses cyclic translations between modalities to generate other modalities only by one modality. Thus, robust joint representations can be learned. T2FN [12] achieve better performance in missing tasks by supervising the learning of representations under the tensor rank regularization. However, T2FN needs aligned inputs. Our proposed model can be treated as an MSA model for wilder application scenarios, which are closer to real circumstances. When there is no missing for any modal, the proposed model works like other conventional MSA models.

2.2 Transformer and BERT

Transformer [23] is a sequence-to-sequence model which is usually used in machine translation tasks. Its attention mechanism provides an effective tool for extracting contextual information from any sequence. As for the MSA task, MulT [22] uses a transformer-based structure to capture the connection between any two modalities. BERT [7] is a large pre-trained language model, which uses a transformer encoder as the basic unit. BERT has prominent results on most NLP tasks. Our proposed model also uses BERT as an effective text feature extraction method and uses transformer encoders to effectively capture the intra- and inter-modal interactions.

2.3 Generative Networks

The generation networks learn the joint probability distribution of the sample and label through the training data. Thus, the trained model can generate new data in line with the sample distribution. The typical generative networks include generative adversarial network(GAN) [8] and variational auto encoder [11]. For modality missing problem, a group of methods [2, 3, 19, 26] utilize GAN or its varieties including cGAN [15] and cycleGAN [33] to generate the data of missing modalities. The CRA [21] use cascaded residual autoencoder adapted from stacked denoising autoencoder [24] to calculate the residual and reconstruct the corrupted multimodal data sequence.

However, the methods based on generative models usually have narrow applications where only one specific modality of samples is missed because one generator can only generate one specific modality from another specific modality. Our proposed model is also different from the autoencoders that aim to impute the complete samples for downstream works. Compared to the generation of missing data, a better feature extraction method also counts in the MSA task. Our proposed model works like a denoising autoencoder whose structure is like CRA but the decoder carries out the task of supervising the learning of effective representations, while the final purpose is still the sentiment prediction.

3 METHODOLOGY

In this section, we describe our approach for learning robust representations against missing modalities through modality reconstruction. The TFR-Net can be segmented into three sub-modules:

modality feature extraction module(Section 3.2), modality reconstruction module (Section 3.3), and fusion module (Section 3.4). The overall framework is illustrated in Fig 2.

3.1 Task Setup

Our goal is to judge the sentiments in videos by leveraging incomplete multimodal signals. For each video clips, three sequences of low-level features with random missing from text (t), audio(a), visual (v) are involved. These are represented as $U_t^{'} \in R^{T_t \times d_t}$, $U_a^{'} \in R^{T_a \times d_a}$, $U_v^{'} \in R^{T_v \times d_v}$ respectively. Proposed model takes $U_t^{'}$, $U_a^{'}, U_v^{'}$ as inputs and outputs one sentimental intensity result $\hat{y}_m$. Besides, for the training stage, complete modality feature $U_t \in R^{T_t \times d_t}$, $U_a \in R^{T_a \times d_a}$, $U_v \in R^{T_v \times d_v}$ and feature missing position are used to guide representation learning.

3.2 Modality Feature Extraction Module

The modality feature extraction module first processes the incomplete modality sequences with a 1D convolutional layer to ensure each element of the input sequences aware of its neighbor elements.

$$H_m = \mathrm{Conv1d}(U_m^{'}, k_m) \in R^{T_m, d}, m \in \{t, a, v\}, \tag{1}$$

where $k_{t,a,v}$ are the sizes of the convolutional kernels for modalities t, a, v, and d is a common dimension. We then augment the convolved sequences with position embedding (PE), followed by intra-modal and inter-modal transformers to capture modality dynamics for each time-step of the input sequences. Utilizing the attention mechanism to extract information for one sequence H_i from another sequence H_j, the transformer encoder structure is used for those transformers. Queries, keys, and values are inputs for a transformer encoder. The source of queries is from H_i while the source of keys and values should be from H_j. So the transformer encoder can be denoted as $\mathbf{Transformer}(H_i, H_j, H_j)$.

$$H_m^{'} = H_m + \mathrm{PE}_m(T_m, d) \tag{2}$$

$$H_{m \to m} = \mathbf{Transformer}\left(H_m^{'}, H_m^{'}, H_m^{'}\right) \in R^{T_m, d} \tag{3}$$

$$H_{n \to m} = \mathbf{Transformer}\left(H_m^{'}, H_n^{'}, H_n^{'}\right) \in R^{T_m, d}, \tag{4}$$

where $\mathrm{PE}_m(T_m, d) \in R^{T_m, d}$ computes the embeddings for each position index, $m \in \{t, a, v\}, n \in \{t, a, v\} - \{m\}$.

Finally, we concatenate all latent features obtained with all intra-modal and inter-modal transformers as the enhanced sequence features output.

$$H_m^{''} = \mathrm{Concat}\left([H_{m \to m}; H_{n_1 \to m}; H_{n_2 \to m}]\right) \in R^{T_m, 3d}, \tag{5}$$

where $m \in \{t, a, v\}$ and n_1, n_2 represent other two modalities except for m. The enhanced sequences are expected to extract the effective representation for the missing modality features taking advantage of the complementarity between modalities. Moreover, such enhanced modality sequences containing cross-modal interactions can be regarded as a model-level fusion result.

3.3 Modality Reconstruction Module

We propose a Modality Reconstruction (MR) Module based on the key insight that reconstructing complete modality sequences from extracted modality sequences can lead the extractor module to learn the semantics of the missing parts. For each modality, a self-attention mechanism on feature dimension is first conducted to capture the interactions among extracted features.

$$H_m^{*} = \mathbf{Transformer}\left(H_m^{''T}, H_m^{''T}, H_m^{''T}\right)^T \in R^{T_m, 3d}, \tag{6}$$

where $m \in \{t, a, v\}$, and H_m^{*} is regraded as transformed sequence features. Then we perform a linear transformation mapping the extracted features into input spaces.

$$\hat{U}_m = W_m \cdot H_m^{*} + b_m, \tag{7}$$

where $m \in \{t, a, v\}$, and W_m, b_m are the parameters of the linear layer.

For supervisions, SmoothL1Loss($\cdot$) between original and generator on missing elements are utilized as the generation loss $\mathcal{L}_g^m$ to leverage the effect of missing reconstruction.

$$\mathcal{L}_g^m = \mathrm{SmoothL1Loss}\left(\hat{U}_m * M', U_m * M'\right), \tag{8}$$

where $m \in \{t, a, v\}$, M' is the missing mask revealing missing positions in input modality sequences.

3.4 Fusion Module

After enhancing the incomplete modalities sequence with complementary modality information under the guidance of reconstruction loss, we fuse them into a joint vector for sentiment predictions. Proposed CNN Gate Encoder is used to encode enhanced modality sequences $\overline{H}_m$ separately.

CNN Gate Encoder. Firstly, extracted modality sequences $\overline{H}_m$ are processed with a bidirectional GRU layer, followed by the *tanh* activation function to get the updated representation $H_m^{''}$.

$$\overline{H}_m = \tanh(\mathrm{BiGRU}(H_m^{''})) \tag{9}$$

A Convolution Gate component is designed to further encode $\overline{H}_m$. Specifically, a one-dimensional convolution network (CNN) slides a convolution kernel with the window size k over the input $H_m^{''}$ to get a scalar value g_i for each element in the sequences. Padding strategy is used to ensure that $H_m^{''}$ and g have the same sequence length:

$$g = \mathrm{sigmoid}(\mathrm{Conv1d}\left(\overline{H}_m\right)), \tag{10}$$

where $m \in \{t, a, v\}$, and $Conv1d(\cdot)$ is an one-dimensional convolution operation. g is regarded a gate to scale the representation $\overline{H}_m$, filtering out irrelevant contextual information in the utterance:

$$\overline{H}_m^{'} = \overline{H}_m \otimes g, \tag{11}$$

where $\otimes$ means element-wise product.

In addition, the representation $\overline{H}_m^{'}$ and the initial extracted sequences $H_m^{''}$ are concatenated. Then a fully connected layer is used

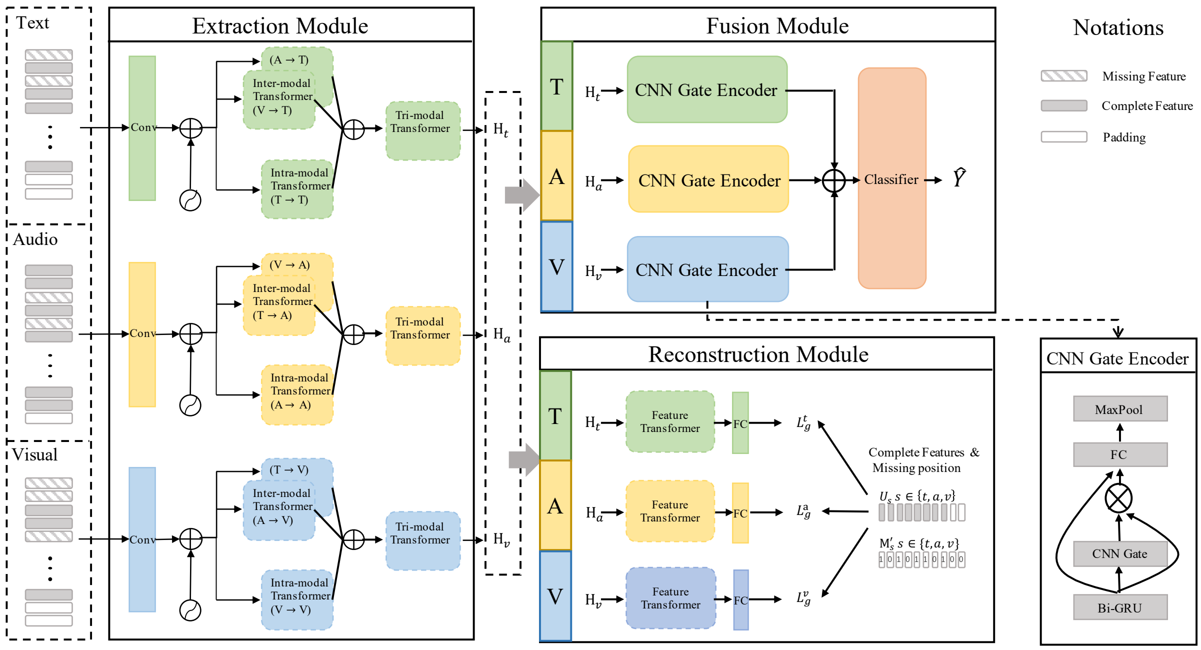

Figure 2: The overall framework of TFR-Net which contains three modules: feature extraction module, modality reconstruction module, and modality fusion module.

to control the final word-level representation H_m^* dimension:

$$H_m^* = \tanh(W \cdot \text{Concat}\left[\overline{H}_m^{'}, H_m^{''}\right] + b) \tag{12}$$

Finally, utilizing the max-pooling operation to focus on features in an utterance that have a more significant impact, the final modality representation U_m^* is defined as follows:

$$U_m^* = \text{Maxpool}\left\{H_m^*\right\} \in R^{h_m}, \tag{13}$$

where h_m means the hidden dimension for modality m.

The concatenation of three modality representation is regarded as the fusion results and is fed into a simple classifier to make a final prediction of the sentiment intensity.

$$U^* = \text{Concat}\left[U_t^*, U_a^*, U_v^*\right] \tag{14}$$

$$\hat{y} = W_1 \cdot \text{LeakyReLU}\left(W_2 \cdot \text{BN}\left(U^*\right) + b_2\right) + b_1, \tag{15}$$

where BN is the BatchNorm operation, and LeakyReLu is used as activation.

3.5 Model Training

We take the L1Loss as the basic optimization objective for sentiment intensity prediction. Along with the reconstruction loss $\mathcal{L}_g^m, m \in \{t, a, v\}$, the overall learning of the model is performed by minimizing:

$$\mathcal{L}_{gen} = \sum_{m \in \{t,a,v\}} \lambda_m \cdot \mathcal{L}_g^m \tag{16}$$

$$\mathcal{L} = \frac{1}{N} \sum_i^N (|\hat{y}^i - y^i|) + \mathcal{L}_{gen} \tag{17}$$

Here, $\lambda_m, m \in \{t, a, v\}$ are the weights that determine the contribution of each modality reconstruction loss $\mathcal{L}_g^m$ to the overall loss $\mathcal{L}$. Each of these component losses are responsible for representation learning in each modality subspace.

Dataset	# Train	# Valid	# Test	# All
MOSI	552/53/679	92/13/124	379/30/277	2199
SIMS	742/207/419	248/69/139	248/69/140	2281

Table 1: Dataset statistics for benchmark MSA dataset in format negative/neutral/positive.

4 EXPERIMENTAL SETUP

In this section, we describe our experimental methodology to evaluate model robustness against random modality feature missing.

4.1 Datasets

In this work, experiments are conducted on two public multimodal sentiment analysis datasets, MOSI [31] and SIMS [28]. The basic statistics of each dataset are shown in Table 1. Here, we give a brief introduction to the above datasets.

MOSI. The CMU-MOSI dataset [31] is one of the most popular benchmark datasets for MSA. The dataset contains 2199 short monologue video clips taken from 93 Youtube movie review videos. The utterances are manually annotated with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

SIMS. The SIMS dataset [28] is a Chinese MSA benchmark dataset with fine-grained uni-modal annotations. The dataset comprises of 2,281 refined video clips collected from different movies, TV serials, and variety shows with spontaneous expressions, various head poses, occlusions, and illuminations. The utterances are manually annotated with a sentiment score from -1 (strongly negative) to 1 (strongly positive).

4.2 Feature Extraction

For each of the three modalities, we process the information from videos as follows.

4.2.1 Text Modality. For both MOSI and SIMS datasets, Pre-trained BERT [7] is utilized as the feature extractor to encode transcribed word sequences into the text modality features U_t with d_t equal to 768.

4.2.2 Audio Modality. For Audio feature extraction, COVAREP acoustic framework [6] is utilized for MOSI dataset, while LibROSA [14] is used for SIMS dataset. The feature dimensions d_a are 5 for MOSI and 33 for SIMS dataset.

4.2.3 Visual Modality. MOSI use Facet[1] to extract facial expression features. For SIMS, MTCNN face detection algorithm [32] is used to extract aligned faces followed by facial features extraction using MultiComp OpenFace2.0 toolkit [1]. The feature dimensions, d_v are 20 for MOSI and 709 for SIMS.

4.3 Baselines

The experiments are conducted on three baseline methods along with our proposed model to validate its performance. All methods can work on unaligned multimodal datasets.

TFN. Tensor Fusion Network (TFN) [29] utilizes tensor fusion layer where a cartesian product is used to form a feature vector. Therefore, information from three modalities can be fused to predict the sentiment.

MulT. Multimodal transformer (MulT) [22] uses a crossmodal attention mechanism to capture the relationship between different modalities. These interactions lead to better performance on unaligned multimodal datasets.

MISA. This method learns both modality-invariant and -specific representations [10] by projecting each modality of samples into two subspaces. This efficient method of feature extraction improves model performance in MSA tasks.

4.4 Experimental Settings

The multimodal datasets which have missing values are built for the experiments. The random replacement with missing values in the sequence is applied to simulate the scene of incomplete multimodal data. The missing value is [UNK] in data of text modality and zero padding vector in data of other modalities. The missing proportion of each modality needs to be specified in advance, and the proportion is the same among train, validate and test datasets. For our proposed model, the hyper-parameters include convolution kernel size, attention dropout, heads of transformers, dimension of feature vectors for fusion, and weights of generative loss for three modalities. Those parameters are well-tuned for different datasets on the valid set. The Adam optimizer is used for learning with the learning rate is 0.002 in the MOSI dataset, 0.001 in the SIMS dataset. The evaluating indicators in results are average of three experiments using three different random seeds in the MOSI and SIMS dataset.

[1]https://imotions.com/platform/

4.5 Evaluation Metrics

For a comprehensive comparison with baselines, binary classification accuracy (Acc-2), five classification accuracy (Acc-5), Mean Absolute Error (MAE), and Pearson Correlation coefficient (Corr) on MOSI and SIMS test sets are recorded with respect to the increasing missing rate. Following the recent works [10, 22], binary classification accuracy (Acc-2) is calculated with the more accurate formulation of negative/positive classes where negative and positive classes are assigned for < 0 and > 0 sentiment scores, respectively. Besides, we compute the Area Under Indicators Line Chart (AUILC) value for each metric sequence to evaluate the overall performance of dealing with incomplete modality input quantitatively. AUILC value is defined as follow:

Area Under Indicators Line Chart (AUILC) Given the model evaluation results sequence $X = \{x_0, x_1, \cdots, x_t\}$ with increasing missing rates $\{r_0, r_1, \cdots, r_t\}$, the Area Under Indicators Line Chart (AUILC) is defined as:

$$\text{AUILC}_{\text{X}} = \sum_{i=0}^{t-1} \frac{(x_i + x_{i+1})}{2} \cdot (r_{i+1} - r_i) \tag{18}$$

For all the above metrics, higher values indicate stronger performance, except MAE where lower values indicate stronger performance.

5 RESULTS AND DISCUSSION

This section presents a detail analysis and discussion about our experimental results.

5.1 Model Robustness for Various Missing Rates

We first study the robustness of the TFR-Net under increasing random modality missing rate. Same level missing rate is introduced in each modalities during both training and testing periods parametrized by **missing_rate** $\in \{0.0, 0.1, \cdots, 1.0\}$. Random drop strategy is utilized for the following experiments, which means each entry is dropped independently with probability $p \in missing\_rate$. Detailed missing construction approach has been described in Section 4.4.

Question1: How does TFR-Net perform compared with existing approaches for multimodal sentiment analysis?

The model performance curve is first illustrated to evaluate model effectiveness intuitively. As shown in Figure 3, on MOSI dataset, TFR-Net surpass baseline approaches on most evaluation metrics for all missing rate $p \in \{0.0, 0.1 \cdots, 1.0\}$. While on SIMS dataset, as shown in Figure 4, TFR-Net achieves better performance under low missing rate ($p \in \{0.0, 0.1, \cdots, 0.5\}$). Under a higher missing rate ($p \in \{0.6, 0.7, \cdots, 1.0\}$), all models perform similarly and finally converge to a stable value. We attribute this phenomenon to the label bias in the dataset. According to the statistics of the positive and negative samples of each dataset in Table 1, we find that there is an obvious label bias on the SIMS dataset. Resulting from the label bias, a trivial model that makes predictions with the average sentiment intensity in the valid set performs well enough. With the

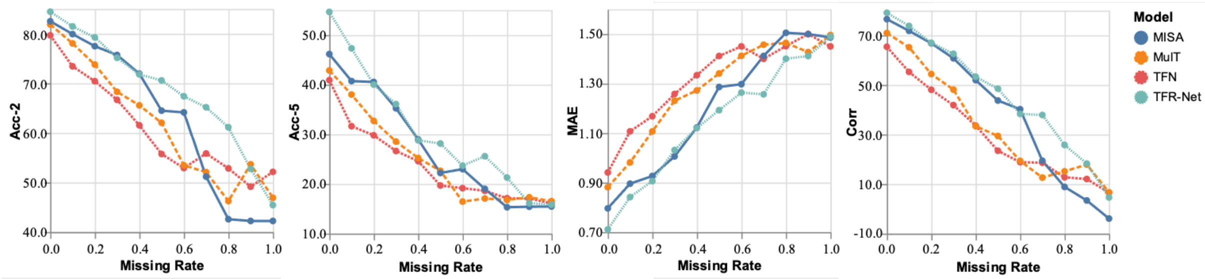

Figure 3: Metrics curves of various missing rates on MOSI dataset.

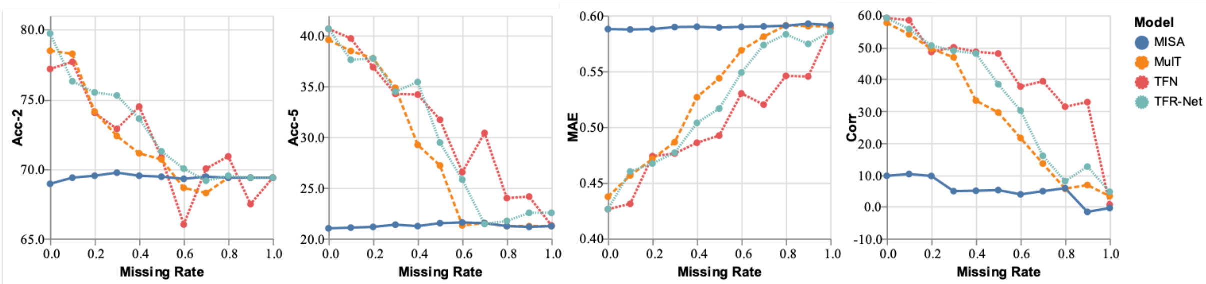

Figure 4: Metrics curves of various missing rates on SIMS dataset.

Models	MOSI				SIMS			
	Acc-2(↑)	Acc-5(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	Acc-5(↑)	MAE(↓)	Corr(↑)
TFN	0.604	0.233	1.327	0.300	0.373	**0.181**	**0.233**	**0.259**
MulT	0.618	0.244	1.288	0.334	0.370	0.173	0.244	0.227
MISA	0.632	0.271	1.209	0.403	0.347	0.106	0.294	0.038
TFR-Net	**0.690**	**0.304**	**1.155**	**0.467**	**0.377**	0.180	0.237	0.253

Table 2: AUILC results comparison with baseline models on MOSI and SIMS dataset. Results on MOSI dataset is calculated with the whole missing rate interval $p \in \{0.0, 0.1, \cdots, 1.0\}$, while results on SIMS dataset is calculated with partial missing rate interval $p \in \{0.0, 0.1, \cdots, 0.5\}$.

missing rate increasing, models struggle to beat such models with information-less train data, and finally degenerate to the trivial ones. Besides the line charts, the AUILC value is calculated to evaluate the proposed model quantitatively. We record the AUILC value of the whole interval $p \in \{0.0, 0.1 \cdots, 1.0\}$ on MOSI dataset, and the AUILC value of partial interval $p \in \{0.0, 0.1 \cdots, 0.5\}$ on SIMS dataset as the indistinguishable results for higher p. From Table 2, the quantitative results further verify the proposed TFR-Net robustness for various modality missing rates on both datasets.

5.2 Model Robustness for Modality Missing Combinations

Our next experiment focuses on the robustness of the TFR-Net for different modality missing combinations. We conduct experiments on the MOSI dataset with TFR-Net under cases where different modality combinations are completely dropped (**missing_rate** $p = 1.0$).

Question2: How does TFR-Net perform with different modality missing combinations during testing?

Experimental results are collected in Table 3. For uni-modal input experiments, we see that TFR-Net maintains comparable performance, while TFR-Net with audio input and visual input fails. For bi-modal input, text modality along with visual modality performs best, and even achieve better MAE and Corr compared with tri-modal inputs. According to the above results, we can summarize that text modality contains more semantics and plays an important role in missing semantics reconstruction and sentiment prediction. While it is relatively hard for the model to reconstruct semantics existing in text modality with audio and visual features input.

5.3 Ablation Study

Finally, the ablation experiment is conducted on the MOSI dataset. We test the intra-modal attention module, generation module, CNN gate module contribution separately. Model without intra-modal is denoted with w/o a, Model without generation module is denote

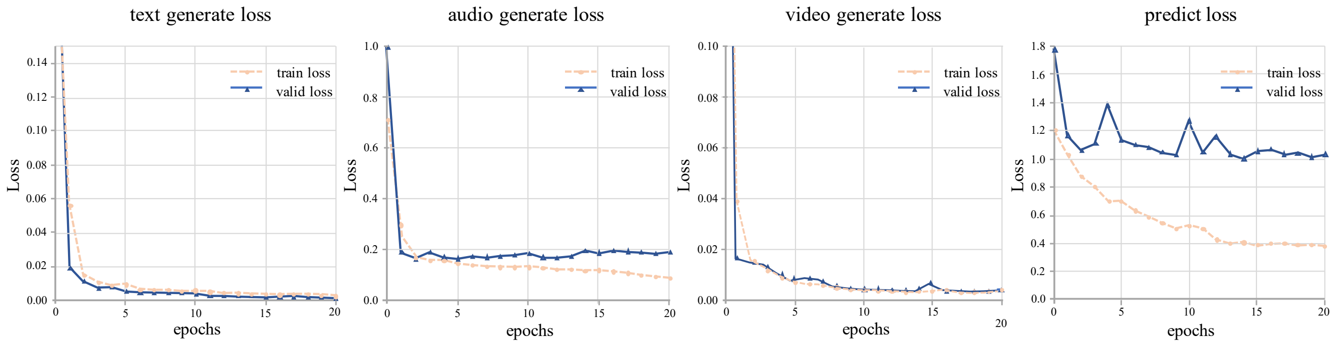

Figure 5: Trends in loss functions in training on MOSI dataset with missing rate of 0.3.

Test Input	MOSI			
	Acc-2(↑)	Acc-5(↑)	MAE(↓)	Corr(↑)
$\{a\}$	55.15	16.57	1.419	0.214
$\{v\}$	60.11	17.49	1.381	0.164
$\{t\}$	83.49	50.14	0.786	0.778
$\{a, v\}$	62.65	19.05	1.334	0.231
$\{t, a\}$	83.99	52.92	**0.731**	**0.788**
$\{t, v\}$	82.62	49.37	0.772	0.778
$\{t, a, v\}$	**84.10**	**54.66**	0.754	0.783

Table 3: TFR-Net results for different modality missing combinations. All results are the average of three groups of seeds.

Metrics (AUILC-)	MOSI			
	Acc-2(↑)	Acc-5(↑)	MAE(↓)	Corr(↑)
TFR-Net (w/o a)	0.671	0.292	1.175	0.461
TFR-Net (w/o g)	0.682	0.301	1.231	0.455
TFR-Net (w/o c)	0.675	0.295	1.167	0.462
TFR-Net	**0.690**	**0.304**	**1.155**	**0.467**

Table 4: Ablation Study on MOSI dataset, (w/o a) means removal of intra-modal attention. (w/o g) means removal of generation module, (w/o c) means removal of CNN gate module.

by w/o g, and Model without CNN gate module is denoted by w/o c. From Table 4, we can see that the removal of any module in TFR-Net results in a decline in model performance. Specifically, the removal of the intra-modal attention module has the most influence on the model performance and leads to a 2% drop in the AUILC value of binary classification accuracy. While the generation module as additional supervision has a relatively small influence on the model performance. To further analyze the effectiveness of the generation module, we illustrate the trends of the generative loss along with prediction loss in both the training and validation period. Figure 5 displays the trend of the SmoothL1Loss of generated three modalities and the regression loss for sentiment prediction. The loss values are traced during the training of TFR-Net on the MOSI dataset, with a missing rate of (0.3, 0.3, 0.3) for three modalities. As shown in Figure 5, the loss values keep descending trend on both training and validation set in the whole training process. Generative loss and prediction loss can converge together. This proves that the model can achieve better sentiment analysis results while learning the representations that can reconstruct the complete multimodal data features.

The above ablation studies' results verify the effectiveness of the proposed module for improving the model robustness against the modality missing.

6 CONCLUSION

In this paper, we stress improving model robustness against incompleteness of modalities for MSA task and design transformer-based feature reconstruction network (TFR-Net), a general framework which is flexible to deal with the incompleteness of non-aligned features in various modality combinations and various degree. At the heart of TFR-Net is the feature reconstruction module, which guides the extractor acquiring semantics of the missing modalities features. All experimental results on two benchmark MSA datasets show that our model achieves good results with the incompleteness of non-aligned features in various modalities and various degrees.

We also find that current model performance is limited by the label bias problem. In future work, we will introduce data augment method dealing with the data bias and further innovate our model with real-time user-generated video input.

ACKNOWLEDGMENTS

This paper is founded by National Key R&D Program Projects of China (Grant No: 2018YFC1707605).

REFERENCES

[1] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 59–66.

[2] Benjamin Bischke, Patrick Helber, Florian Koenig, Damian Borth, and Andreas Dengel. 2018. Overcoming Missing and Incomplete Modalities with Generative Adversarial Networks for Building Footprint Segmentation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*. 1–6.

[3] 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*. 1158–1166.

[4] Chongqing Chen, Dezhi Han, and Jun Wang. 2020. Multimodal encoder-decoder attention networks for visual question answering. *IEEE Access* 8 (2020), 35662–35671.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, Vol. 27. 2672–2680.

[9] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access* 7 (2019), 63373–63394.

[10] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. *CoRR* abs/2005.03545 (2020). arXiv:2005.03545 https://arxiv.org/abs/2005.03545

[11] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR 2014 : International Conference on Learning Representations (ICLR) 2014*.

[12] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1569–1576.

[13] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 2247–2256.

[14] 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 *Proceedings of the 14th python in science conference*, Vol. 8. Citeseer, 18–25.

[15] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784* (2014).

[16] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 973–982.

[17] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6892–6899.

[18] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. 2016. Multimodal video description. In *Proceedings of the 24th ACM international conference on Multimedia*. 1092–1096.

[19] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. 2017. VIGAN: Missing view imputation with generative adversarial networks. In *2017 IEEE International Conference on Big Data (Big Data)*, Vol. 2017. 766–775.

[20] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14. https://doi.org/10.1016/j.imavis.2017.08.003 Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

[21] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4971–4980.

[22] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Vol. 2019. 6558–6569.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Vol. 30. 5998–6008.

[24] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research* 11, 110 (2010), 3371–3408.

[25] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[26] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. 2018. Partial Multi-view Clustering via Consistent GAN. In *2018 IEEE International Conference on Data Mining (ICDM)*. 1290–1295.

[27] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. 11–19.

[28] 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*. 3718–3727.

[29] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[30] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning.. In *AAAI*. 5634–5641.

[31] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[32] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2242–2251.