TMMDA: A New Token Mixup Multimodal Data Augmentation for Multimodal Sentiment Analysis

Xianbing Zhao*
Harbin Institute of Technology
(Shenzhen), Peng Cheng Laboratory
Shenzhen, China
zhaoxianbing_hitsz@163.com

Yixin Chen
Harbin Institute of Technology
(Shenzhen)
Shenzhen, China
cyxhelloo@gmail.com

Sicen Liu
Harbin Institute of Technology
(Shenzhen), Peng Cheng Laboratory
Shenzhen, China
liusicen_cs@outlook.com

Xuan Zang
Harbin Institute of Technology
(Shenzhen)
Shenzhen, China
zangxuan96@gmail.com

Yang Xiang
Peng Cheng Laboratory
Shenzhen, China
xiangy@pcl.ac.cn

Buzhou Tang
Harbin Institute of Technology
(Shenzhen), Peng Cheng Laboratory
Shenzhen, China
tangbuzhou@hit.edu.cn

ABSTRACT

Existing methods for Multimodal Sentiment Analysis (MSA) mainly focus on integrating multimodal data effectively on limited multimodal data. Learning more informative multimodal representation often relies on large-scale labeled datasets, which are difficult and unrealistic to obtain. To learn informative multimodal representation on limited labeled datasets as more as possible, we proposed TMMDA for MSA, a new **T**oken **M**ixup **M**ultimodal **D**ata **A**ugmentation, which first generates new virtual modalities from the mixed token-level representation of raw modalities, and then enhances the representation of raw modalities by utilizing the representation of the generated virtual modalities. To preserve semantics during virtual modality generation, we propose a novel cross-modal token mixup strategy based on the generative adversarial network. Extensive experiments on two benchmark datasets, i.e., CMU-MOSI and CMU-MOSEI, verify the superiority of our model compared with several state-of-the-art baselines. The code is available at https://github.com/xiaobaicaihhh/TMMDA.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Social networks**; • **Computing methodologies** → Artificial intelligence;

KEYWORDS

Multimodal Sentiment Analysis, Data Augmentation, Generative Adversarial Network, Mixup.

ACM Reference Format:
Xianbing Zhao, Yixin Chen, Sicen Liu, Xuan Zang, Yang Xiang, and Buzhou Tang. 2023. TMMDA: A New Token Mixup Multimodal Data Augmentation for Multimodal Sentiment Analysis. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583406

1 INTRODUCTION

With the explosion of video clips in social media, it is critical that deep learning models can accurately predict sentiments in video clips, while language, visual, and acoustic are the three fundamental modalities exhibiting video clips. Multimodal Sentiment Analysis (MSA) employs high-dimensional inputs from modalities as diverse as language, vision, and acoustic to predict sentiment polarity of video clip. Multimodal sentiment analysis, a fundamental and crucial task in affective compute, has become an important area of multimodal research in recent years [23, 27, 44, 48]. Several fusion methods have been proposed for multimodal sentiment analysis[17, 18, 23, 27, 35, 41, 42] and considerable progress has been made. A few existing works [10, 40] have attempted to model heterogeneous multimodal data representation with considering the modalities gap. These approaches have achieved efficient performance, however, these methods only learn multimodal representations with limited annotated data to perform the training of a deep learning model.

Data augmentation training strategy has achieved great success to improve model performance in multiple computer vision (CV) [26, 49] and natural language processing (NLP) [36, 37] tasks. However, it is not straightforward to apply previous data augmentation methods for multimodal sentiment analysis tasks. In addition, these methods have not been explored in the multimodal sentiment analysis tasks.

To address the above challenges, we propose TMMDA, a new token mixup multimodal data augmentation for multimodal sentiment analysis. Inspired by recent studies on some data augmentation techniques [2, 9], cross-lingual tasks [4, 16], and visual tasks [24, 34], we employ the complementary nature of multimodal data to perform data augmentation. As shown in Fig. 1, previous works directly fuse limited multimodal data to make sentiment predictions. Unlike these methods, our approache generates new text, visual, and

[*] Buzhou Tang is corresponding author. This paper was completed during the internship in Peng Cheng Laboratory.

WWW '23, April 30–May 04, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583406

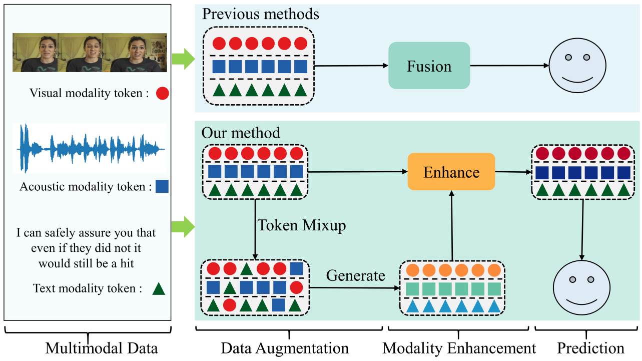

Figure 1: An illustration of our proposed TMMDA. TMMDA explores data augmentation technique and how augmented data can be used to improve accuracy.

acoustic modalities with semantic relationships preserved by token mixup and generative adversarial network. Immediately, we utilize the generated virtual modalities to enhance the corresponding raw modalities and make sentiment predictions. Concretely, we first introduce a cross-modal token mixup (CTM) module. CTM generates a new training sample by constructing manifold mixup interpolations, and keep the semantics of mixed sequence unchanged. Each new training sample contains tokens from three modalities. The mixed sequences still retain the semantic information [4] of the raw modalities, hence we utilize the augmented data to improve model training. Afterward, we design a cross-modal generative adversarial network (CGAN) for each modality. The generator of CGAN takes a random noise sampled from a Gaussian distribution and the new training sample as inputs. We steer the generator network to produce a virtual modality representation, and the discriminator to reduce the semantic gap of raw modality and generated virtual modality. In addition, we minimize the Jensen−Shannon divergence discrepancy distance between the representations of generated virtual modality and raw modality, which encourages the model to learn underlying semantic similarities. Meanwhile, we also elaborate a cross-modal encoder (CME) module to enhance the raw modality, which dynamically determines the passed proportions of the generated virtual modality information. Extensive experimental results on two benchmark datasets CMU-MOSI and CMU-MOSEI, validate the effectiveness and superiority of our proposed method. The main contributions of this paper are three-fold:

- We introduce a cross-modal token mixup module, which creates new training samples by constructing manifold mixup interpolations, and preserves the semantic relationship of raw modality.
- We propose a cross-modal adversarial training strategy to generate new virtual modalities by using the new training samples. We design a cross-modal encoder module to dynamically filter the information of generated virtual modality related to the raw modality and enhance the raw modality.
- Extensive experiments on CMU-MOSI and CMU-MOSEI datasets demonstrate that TMMDA outperforms state-of-the-art methods.

2 RELATED WORK

2.1 Multimodal Sentiment Analysis

According to the difference of core idea in learning unified representations, we roughly divide existing methods into two categories: through loss back-propagation or geometric manipulation in the feature spaces

The former accomplishes multimodal fusion from tensor fusion [17, 41, 43] to an attention-based cross-modal interaction method [18, 23, 27, 48] with passing the task loss through backpropagation. Concretely, Zadeh et al. [41] constructed a Tensor Fusion Network to learn intra-modality and inter-modality dynamics end-to-end. Liu et al. [17] proposed Lowrank Multimodal Fusion method to improve efficiency by performing multimodal fusion using low-rank tensors. Zadeh et al. [42] designed a Memory Fusion Network which explicitly accounts for both interactions in a neural architecture and continuously models them through time. Wang et al. [35] proposed a Recurrent Attended Variation Embedding Network considers the fine-grained structure of nonverbal subword sequences and dynamically shifts the word representations based on these nonverbal cues. Tsai et al. [27] proposed Multimodal Transformer which extends the standard Transformer network to learn representations directly from unaligned multimodal streams. Rahman et al. [23] designed a Multimodal Adaptive Gate to integrate multimodal information into pre-trained language representation model by employing multimodal adaptive gate dynamically filter information. Lv et al. [18] proposed Progressive Modality Reinforcement method to explore the three-way interactions across all the involved modalities under the background of multimodal fusion from unaligned multimodal. yang et al. [38] proposed a ModalTemporal Attention Graph which provides a suitable framework for analyzing multimodal sequential data by employing interpretable graph-based neural model. Zeng et al. [45] proposed a Modulation Model to identify the contribution of modalities and reduce the impact of noisy information. Zhao et al. [48] proposed a MAG+ which extends MAG to reinforce multimodal fusion. The latter branch targets at modeling heterogeneous multimodal data via exploiting geometric manipulation in the feature spaces. Sun et al. [25] proposed an Interaction Canonical Correlation Network to explore correlations between all three modes via deep canonical correlation analysis. Hazarika et al. [10] proposed a multimodal framework that learns factorized subspaces for each modality and provides better representations as input to fusion. Yu et al. [40] proposed a self-supervised multi-task learning strategy to acquire independent unimodal supervisions. Han et al. [7] propose a Bi-Bimodal Fusion Network which performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations. Han et al. [8] proposed a MultiModal InfoMax method, which hierarchically maximizes the mutual information in unimodal input pairs and between multimodal fusion result and unimodal input in order to maintain task related information through multimodal fusion.

Although significant performance has been achieved by considering different fine-grained fusion methods and representation learning methods, however, these methods only focus on fusing limited multimodal data. To this end, designing a data augmentation method, which could directly enhance the multimodal representation, becomes a critical challenge.

2.2 Token Mixup

Our work is inspired by the token mixup methods. Zhang et al. [46] first proposed mixup to alleviate that large deep neural networks exhibits undesirable behaviors such as memorization and sensitivity to adversarial examples. Verma et al. [31] proposed manifold mixup which employs semantic interpolations as enhanced training information to obtain model with smoother decision boundaries at fine-gained representation. Recent studies have introduced mixup on multiple tasks such as multimodal vision [15, 24, 34], machine translation [3, 6, 14], and speech recognition [4, 19]. and has been made considerable progress. Our approach is the first to introduce the idea of mixup to perform data augmentation for multimodal sentiment analysis tasks.

2.3 Generative Adversarial Network

Generative Adversarial Network (GAN) was first introduced by Goodfellow et al. [5] in 2014. GAN has made a considerable progress in multiple tasks, such as computer vision [1, 21, 32, 33] and natural language processing [12, 13, 39, 47], among others, due to its generative capabilities to generate realistic examples plausibly drawn from an training data distribution. We employ generative adversarial network to generate a new modality and enhance raw modality.

3 METHODOLOGY

3.1 Problem Statement

The goal of multimodal sentiment analysis is to employ multimodal data, i.e., text, visual, and acoustic modalities to make sentiment predictions. The three input sequences of the aforementioned modalities are denoted by $X_L \in \mathbb{R}^{T_l \times d_l}$, $X_A \in \mathbb{R}^{T_a \times d_a}$, and $X_V \in \mathbb{R}^{T_v \times d_v}$, respectively. T_* and $d_*, * \in \{l, v, a\}$ represent the sequence length and feature dimension, respectively. Our goal is to enhance multimodal representations via data augmentation technique, and produce desirable performance for multimodal sentiment analysis.

3.2 Overall Architecture

Our model is trained in an end-to-end manner. We first employ a fully connected layer to process the raw modalities and unify the feature dimensions of different modalities, which are denoted by $X_L \in \mathbb{R}^{T_l \times d}$, $X_A \in \mathbb{R}^{T_a \times d}$, and $X_V \in \mathbb{R}^{T_v \times d}$. The raw feature of visual and acoustic modalities are pre-extracted from CMU-MOSI [43] and CMU-MOSEI [44] datasets, respectively. Text features are extracted by the pre-trained language representation model, i.e., BERT [11]. To capture the sequence-level context of visual and acoustic modalities, we use a 1-layer Transformer encoder [30] to model multimodal long-term contextual information. For each modality (taking text modality as an example), we mix the token of text and the other two modalities (i.e., visual and acoustic) to create a new training sample, i.e., mixed sequence. Immediately, the text modality and mixed sequence pass through a cross-modal generative adversarial network, which makes discrimination in both the text representation and the generated virtual text modality mutually boosting. Adversarial training correlates text representation and virtual text representation. In addition, we adopt Jensen-Shannon divergence to further reduce modality gap and help fusion. Immediately, we use the generated text representation to enhance text representation via cross-modal encoder. Finally, a 1-layer Transformer encoder and several fully-connected layers are designed to make sentiment predictions. Fig. 2 shows the information flow across the main framework of TMMDA, which comprises the following modules: Cross-modal Token Mixup, Cross-modal Generative Adversarial Network, and Cross-modal Encoder.

3.3 Cross-modal Token Mixup

In this section, we introduce the Cross-modal Token Mixup (CTM), a multi-modal joint data augmentation, to mix up the tokens of text, visual, and acoustic modality by controlling multiple mixup ratios. Note that we perform mixup using align datasets, where the pre-extracted are performed a word-level forced alignment [23] among text transcriptions, visual, and acoustic representations. As shown in Fig. 1, CTM creates three new training samples (i,e., $\dot{X}_{LM}$, $\dot{X}_{VM}$, and $\dot{X}_{AM}$) by employing linearly interpolating text, visual, and acoustic sequences. CTM receives three input sequences, which are purely text sequence X_L, visual sequence X_V, and acoustic sequence X_A.

$$X_L = \{x_{l_1}, x_{l_2}, ..., x_{l_n}\} \tag{1}$$

$$X_V = \{x_{v_1}, x_{v_2}, ..., x_{v_n}\} \tag{2}$$

$$X_A = \{x_{a_1}, x_{a_2}, ..., x_{a_n}\} \tag{3}$$

where subscript n denotes the length of sequence.

Given a triplet $(x_{l_i}, x_{v_i}, x_{a_i})$, we perform cross-modal token mixup for text sequence X_L, visual sequence X_V, and acoustic sequence X_A, respectively. To create a mixed sequence of text modality, for each generated token X_{lmi}, we choose a token among three input sequences X_L, X_V, and X_A with three certain probabilities ρ_l, ρ_{la}, and ρ_{lv}. The subscript m denotes mixed sequence, and the subscript i denotes the i-th token of the sequence. The new training sequence of text modality is created via:

$$x_{lm_i} = \begin{cases} x_{v_i}, 0 \le \rho_l \le \rho_{lv} \\ x_{l_i}, \rho_{lv} < \rho_l < \rho_{la} \\ x_{a_i}, \rho_{la} \le \rho_l \le 1 \end{cases} , \rho_l \sim U(0,1) \tag{4}$$

where ρ_l is sampled from the uniform distribution $U(0, 1)$, and ρ_{lv} and ρ_{la} are two hyper-parameters between 0 and 1. The x_{lm_i} denotes the i-th token of new training sequence. Finally, we concatenate all the token x_{lm_i} together and obtain the corresponding mixed sequence $\dot{X}_{LM}$ of text modality:

$$\dot{X}_{LM} = \{x_{lm_1}, x_{lm_2}, ..., x_{lm_n}\} \tag{5}$$

Similarly, we can obtain new training sequences of visual and acoustic modality by employ CTM, respectively. the newly generated training sequences $\dot{X}_{AM}$ and $\dot{X}_{VM}$ in its embedding form is:

$$x_{vm_i} = \begin{cases} x_{l_i}, 0 \le \rho_v \le \rho_{vl} \\ x_{v_i}, \rho_{vl} < \rho_v < \rho_{va} \\ x_{a_i}, \rho_{va} \le \rho_v \le 1 \end{cases} , \rho_v \sim U(0,1) \tag{6}$$

$$\dot{X}_{VM} = \{x_{vm_1}, x_{vm_2}, ..., x_{vm_n}\} \tag{7}$$

$$x_{am_i} = \begin{cases} x_{l_i}, 0 \le \rho_a \le \rho_{al} \\ x_{a_i}, \rho_{al} < \rho_a < \rho_{av} \\ x_{v_i}, \rho_{av} \le \rho_a \le 1 \end{cases} , \rho_a \sim U(0,1) \tag{8}$$

$$\dot{X}_{AM} = \{x_{am_1}, x_{am_2}, ..., x_{am_n}\} \tag{9}$$

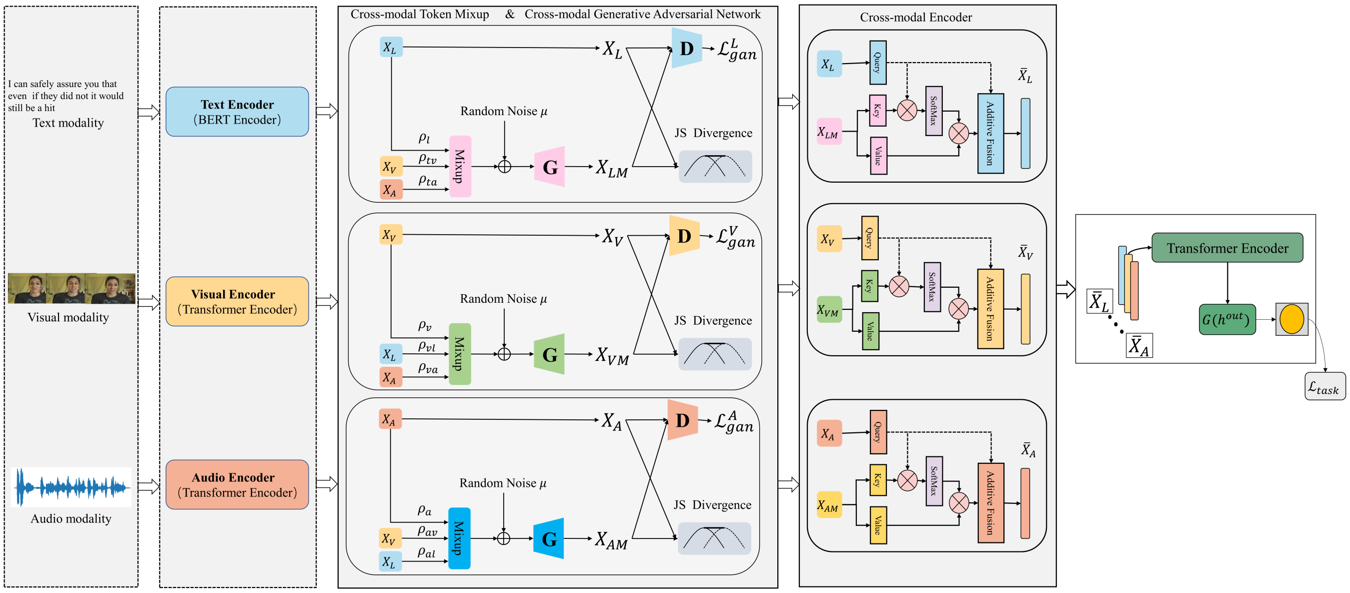

Figure 2: Overall structure of TMMDA.

where the x_{vm_i} and X_{am_i} denotes the i-th token of the generated visual and acoustic sequences, respectively. ρ_v, ρ_a are sampled from the uniform distribution $U(0,1)$. We use these new samples to help adversarial training.

3.4 Cross-modal Generative Adversarial Network

The basis Generative Adversarial Network (GAN) consists of a generative model G and a discriminative model D, where G capture the distribution over training data and D that distinguishes between synthesize and real samples [5]. As shown in Fig. 1, the CGAN consist of three parallel generative and discriminative networks for text, visual, and acoustic modalities, respectively. Both generative and discriminative networks are composed of several fully connected layers and activation functions. The generative network and discriminative network are trained together to correlate the representation of mixed sequence and raw modality. In addition, we optimize the discrepancy distance objective by enforcing the similarity between the distributions of generated virtual modality representation and raw modality representation in the representation space, which encourages the model to learn cross-modal correlation to reduce the gap between the two representations before fusion. Concretely, the generator take a random noise variant μ and mixed sequence as its inputs, where μ is sampled from a Gaussian distribution $\mathcal{N}(0,1)$. We define a new input transformation:

$$\hat{X}_{LM} = \dot{X}_{LM} + \alpha_l \cdot \mu, \mu \sim \mathcal{N}(0,1) \tag{10}$$

$$\hat{X}_{VM} = \dot{X}_{VM} + \alpha_v \cdot \mu, \mu \sim \mathcal{N}(0,1) \tag{11}$$

$$\hat{X}_{AM} = \dot{X}_{AM} + \alpha_a \cdot \mu, \mu \sim \mathcal{N}(0,1) \tag{12}$$

where α_l, α_v, and α_a are hyperparameters to regulate the random noise ratio, respectively. The generators of CGAN take $\dot{X}_{*M}$ as inputs and output generated virtual modalities $X_{*M}, * \in \{L, A, V\}$. The objective function is given as follows:

$$\mathcal{L}^*_{\text{gan}} = \mathbb{E}_{p(x_*),p(x_{*m})}\left[\log D^*(x_*) + \log\left(1 - D^*(x_{*m})\right)\right], \quad * \in \{L, V, A\} \tag{13}$$

where $p(x_*)$ and $p(x_{*m})$ denote the distribution of raw modalities X_* and generated virtual modalities X_{*M}. To further reduce the gap between the generated virtual modality and raw modality, we employ Jensen–Shannon divergence (JSD) to enforce the similarity between the distribution of the raw modalities and generated virtual modalities in the representation space. Specifically, The JSD is defined as:

$$D_{KL}(p\|q) = -\sum_x p(x)\log\frac{q(x)}{p(x)} \tag{14}$$

$$M = \frac{1}{2}(P+Q) \tag{15}$$

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M) \tag{16}$$

$$\mathcal{L}^*_{dist} = JSD(P(x_*)\|P(x_{*m})), * \in \{L, V, A\} \tag{17}$$

where D_{KL} is the Kullback-Leibler divergence function. $P(x_*)$ and $P(x_{*m})$ are the distribution of X_* and X_{*M}, respectively. X_{*M} represents the representation of generated virtual modality.

3.5 Cross-modal Encoder

Given two sequences $X_{S1} \in \mathbb{R}^{T_{s1}\times d}$ and $X_{S2} \in \mathbb{R}^{T_{s2}\times d}$, where T and d denote sequence length and feature dimension, respectively. $S1$ and $S2$ represent source sequence and target sequence, respectively. Based on the attention mechanism [30], we design the Cross-modal Encoder (CME) module that enables modality X_{S1} for receiving filtered information from modality X_{S2}. Generally, cross-modal attention is limited by the feature distribution of various modalities that are different due to heterogeneity, which poses a great

Table 1: Predicted results of TMMDA on datasets CMU-MOSI and CMU-MOSEI. the numbers before '/' denote the average results of 5 runs, while the numbers after '/' denote the best results of 5 runs. (G), (B), and (C) indicate that the text representations are extracted by Glove [22], BERT [11], and COCOLM [20], respectively. '↑' indicates good performance for large values, and '↓' indicates good performance for small values.

Model	MOSI				MOSEI			
	Acc-2 ↑	F1-Score ↑	MAE ↓	CC ↑	Acc-2 ↑	F1-Score ↑	MAE ↓	CC↑
TFN (G) [41]	-/80.8	-/80.7	0.901	0.698	-/82.5	-/82.1	0.593	0.700
LMF (G) [17]	-/82.4	-/82.4	0.917	0.695	-/82.0	-/82.1	0.623	0.677
MFN (G) [42]	77.4/-	77.3/-	0.965	0.632	76.0/-	76.0/-	-	-
RAVEN (G) [35]	78.0/-	76.6/-	0.915	0.691	79.1/-	79.5/-	0.614	0.662
MFM (G) [28]	-/81.7	-/81.6	0.877	0.706	-/84.4	-/84.3	0.568	0.717
MulT (G) [27]	-/83.0	-/82.8	0.871	0.698	-/82.5	-/82.3	0.580	0.703
MISA (B) [10]	81.8/83.4	81.7/83.6	0.783	0.761	83.6/85.5	83.8/85.3	0.555	0.756
MTAG (G) [38]	-/82.3	-/82.1	0.866	0.722	-	-	-	-
PMR (G) [40]	-/83.6	-/83.4	-	-	-/83.3	-/82.6	-	-
ICCN (B) [25]	-/83.07	-/83.02	0.862	0.714	-/84.18	-/84.15	0.565	0.713
Self-MM (B) [40]	84.0/86.0	84.4/85.9	0.713	0.798	82.8/85.2	82.5/85.3	0.530	0.765
M3SA (B) [45]	-/85.70	-/85.60	0.714	0.794	-/85.60	-/85.50	0.587	0.789
MMIM (B) [8]	84.14/86.06	84.00/85.98	0.700	0.800	82.24/85.97	82.66/85.94	**0.526**	0.722
BBFN (B) [7]	-/84.30	-/84.30	0.776	0.755	-/86.20	-/86.10	0.529	0.767
MAG (B) [23]	84.20/86.10	84.10/86.00	0.712	0.796	84.70/-	84.50/-	-	-
TMMDA (C) (ours)	**89.62/90.41**	**89.58/90.38**	**0.593**	**0.870**	**87.15/87.87**	**87.07/87.51**	0.547	**0.823**

challenge to multi-modal fusion. We employ cross-modal attention to capture the correlation between the representations of raw modality and generated virtual modality that is semantically similar representations by exploiting CGAN and JSD to reduce the gap. The information flow from X_{S2} to X_{S1} is presented as the cross-modal attention:

$$\begin{aligned} Y_{S1} &= \mathrm{CM}_{S2\rightarrow S1}\left(X_{S1}, X_{S2}\right) \\ &= \mathrm{softmax}\left(\frac{Q_{S1}K_{S2}^{\top}}{\sqrt{d_k}}\right)V_{S2} \\ &= \mathrm{softmax}\left(\frac{X_{S1}W_{Q_{S1}}W_{K_{S2}}^{\top}X_{S2}^{\top}}{\sqrt{d}}\right)X_{S2}W_{V_{S2}} \end{aligned} \tag{18}$$

$$\bar{X}_{S1} = X_{S1} + Y_{S1} \tag{19}$$

where the Query, Key, and Value are defined as $Q_{S1} = X_{S1}W_{S1}$, $K_{S2} = X_{S2}W_{S2}$, and $V_{S2} = X_{S2}W_{S2}$, respectively. Cross-Modal Encoder can be written as $\bar{X}_{S1} = CME(X_{S1}; X_{S2})$. $\bar{X}_{S1}$ denotes the enhanced representation of raw modality receiving filtered information from the corresponding representation of generated virtual modality by employing CME. The three enhanced modality representations are denoted as following:

$$\bar{X}_{*} = CME(X_{*}, X_{*M}), * \in \{L, V, A\} \tag{20}$$

where $\bar{X}_{*}$ i.e., $\bar{X}_{L}$, $\bar{X}_{V}$, and $\bar{X}_{A}$ represent enhanced text, visual, and acoustic representation, respectively. As a final step, we concatenate the outputs of CME in token dimension, and then input the sequence to 1-layer Multimodal Transformer. Fig. 3 displays the information flow. Eventually, we extract the first element $[CLS]$ of the sequence to pass through fully-connected layers to make sentiment predictions.

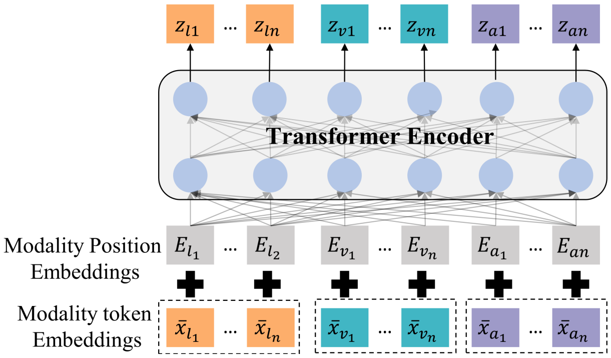

Figure 3: Structure of Multimodal Transformer. Multimodal Transformer takes the sum of modality position embeddings and token embeddings as inputs.

3.6 Objective Formulation

The loss function consists of three parts: task loss, discrepancy distance loss, and adversarial loss. For sentiment intensity prediction, the task loss adopts the mean squared error (MSE) to model the regression problem. For discrepancy distance, we employ Jensen–Shannon divergence loss to facilitate semantically similar representations. For generative adversarial networks, we use the adversarial loss to train the discriminator and generator. These single losses and total loss are calculated as:

$$\mathcal{L}_{task} = \frac{1}{N}\sum_{i=1}^{N}\|y_i - \hat{y}_i\|^2 \tag{21}$$

$$\mathcal{L} = \beta_{dist}\sum_{*\in\{L,V,A\}}\mathcal{L}_{dist}{}^{*} + \beta_{gan}\sum_{*\in\{L,V,A\}}\mathcal{L}_{gan}^{*} + \beta_{task}\mathcal{L}_{task} \tag{22}$$

where β_{dist}, β_{gan}, and β_{task} are tunable parameter to control the power of regularization.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We follow the previous works [23, 27, 40], and conduct experiments on two datasets: CMU-MOSEI [44] and CMU-MOSI [43]. CMU-MOSI is a multimodal sentiment analysis dataset composed of 2199 YouTube video clips. Each multimodal sample has a sentiment score distributed in [-3, 3], where 3 means strongly positive, and -3 means strongly negative. CMU-MOSEI is a dataset of movie reviews collected from YOUTUBE for sentiment analysis. The scoring system of CMU-MOSEI is similar to CMU-MOSI. It contains 22856 video clips. The following four metrics are used to evaluate the performances of all models: Binary Classification Accuracy (Acc-2), F1-Score, Mean Absolute Error (MAE), and Correlation Coefficient (CC). In addition, binary classification accuracy (Acc-2) are calculated by converting the regression output into categorical values. Higher value means better performance for all the metrics except MAE. The above evaluation metrics are consistent with the previous work [23, 27, 40, 48].

Table 2: Performance comparison on CMU-MOSI with different pre-trained language representation model. (B) and (C) indicate BERT [11] and COCOLM [20], respectively. Models with '*' are produced under the same conditions on the CMU-MOSI according to the code provided by the author.

Model	Acc-2	F1-Score	MAE	CC
Visual (Only)	57.40	57.03	1.160	0.143
Audio (Only)	58.17	56.97	1.150	0.144
Text (Only) (B)	84.30	84.30	0.730	0.794
Text (Only) (C)	87.94	87.92	0.708	0.846
MulT* (B)	85.31	85.13	0.734	0.791
MAG (B)	86.10	86.00	0.712	0.796
TMMDA (B)	86.87	86.86	0.703	0.801
MulT* (C)	88.55	88.52	0.654	0.856
MAG* (C)	88.70	88.53	0.624	0.857
TMMDA (C)	**90.41**	**90.38**	**0.593**	**0.870**

4.2 Baselines

Tensor Fusion Network (TFN) [43] employs multimodal tensor to capture inter- and intra-modal interactions. Low-rank Multimodal Fusion (LMF) [17] reduces the computational complexity of multimodal tensors by using low-rank decomposition. Memory Fusion Network (MFN) learns view-specific and cross-view information by using LSTM and memory attention network. Multimodal Factorization Model (MFM) [28] learns multimodal discriminative and modality-specific representations. Multimodal Transformer (MulT) [27] learns cross-modal interactions by using Transformer-based model. Recurrent Attended Variation Embedding Network (RAVEN) [35] employ non-verbal representations to adjust work representations. Interaction Canonical Correlation Network (ICCN) [25] learns multimodal correlation by using deep canonical correlation analysis. Multimodal Adaptation Gate (MAG) [23] integrate non-verbal information into the intermediate layer of pre-trained language model by using adaptive gate. Modality-Invariant and -Specific Representations (MISA) [10] models the modality-specific and modality-invariant representation. Modal-Temporal Attention Graph (MTAG) [38] model heterogeneous unaligned multimodal signals by using an interpretable graph neural network. Progressive Modality Reinforcement (PMR) [18] designs a message center to enhance each modality. Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) [40] introduces a self-supervised method to learn the accurate unimodal label. Modulation Model for Multimodal Sentiment Analysis (M3SA) [45] identifies the contribution of unimodality by according to the noise distribution of each modality. MultiModal InfoMax (MMIM) [8] learns sentiment information by using mutual information of paired modalities. Bi-Bimodal Fusion Network (BBFN) [7] learns relevance and difference of multimodal information.

Table 3: Ablation experiments of TMMDA on the CMU-MOSI dataset. The best results are highlighted in bold.

Model	Acc-2	F1-Score	MAE	CC
Transformer (base)	88.24	88.20	0.671	0.850
w/o CTM	89.31	89.28	0.651	0.852
w/o CGAN	88.55	88.54	0.698	0.851
w/o CME	89.77	89.75	0.597	0.866
w/o JSD	89.92	89.91	0.622	0.859
TMMDA	**90.41**	**90.38**	**0.593**	**0.870**

4.3 Quantitative Analysis

4.3.1 Performance Comparison. To verify the effectiveness of TMMDA, we compare TMMDA with the following state-of-the-art methods: TFN [41], LMF [17], MFN [44], MFM [28], MuLT [27], MISA [10], MTAG [38], PMR [18], MAG-BERT [23] Self-MM [40], M3SA [45], MMIA [8] and BBFN [7]. Table. 1 displays the comparison results. By analyzing this table, we gained the following observations: In terms of cross-modal interaction pattern-based approaches, MAG and MAG+ [23] outperform previous methods with a wide margin by integrating multimodal information in BERT intermediate layers. Concretely, it builds a multimodal adaptive gate to filter non-verbal information with different pre-trained language representation models. This result indicates that elaborately establishing interaction modules and excellent pre-trained language representation models are extremely essential for multimodal sentiment analysis.

Our proposed model TMMDA achieves more prominent performance compared baselines on most criteria of CMU-MOSI and CMU-MOSEI. Compared with MAG and MAG+, our approach obtains relative Acc2 gains with 4.38% and 4.08% on CMU-MOSI, respectively. The improvement indicates the feasibility and importance of powerful enhanced modality representations and advanced pre-trained language representation models. In addition, we also reproduced some state-of-the-art methods and conducted a series of experiments by applying BERT and COCOLM as the pre-trained language representation model, respectively. Table. 2 shows the unimodal baselines and state-of-the-art baselines with different pre-trained language representation models. TMMDA still achieves

Table 4: Comparison of different mixp variants on CMU-MOSI by controlling the condition of mixup ratio.

Mixup Variant (L)	Text		Mixup Variant (V)	Visual		Mixup Variant (A)	Acoustic		Result	
	ρ_{lv}	ρ_{la}		ρ_{al}	ρ_{av}		ρ_{vl}	ρva	Acc-2	F1-Score
$L \rightarrow L$	0.00	1.00	$V \rightarrow V$	0.00	1.00	$A \rightarrow A$	0.00	1.00	88.85	88.82
$V \rightarrow L$	1.00	1.00	$L \rightarrow V$	1.00	1.00	$L \rightarrow A$	1.00	1.00	88.69	88.61
$A \rightarrow L$	0.00	0.00	$A \rightarrow V$	0.00	0.00	$V \rightarrow A$	0.00	0.00	88.70	88.67
$(L, V) \rightarrow L$	0.20	1.00	$(V, L) \rightarrow V$	0.20	1.00	$(A, L) \rightarrow A$	0.20	1.00	89.45	89.43
$(L, V) \rightarrow L$	0.50	1.00	$(V, L) \rightarrow V$	0.50	1.00	$(A, L) \rightarrow A$	0.50	1.00	89.47	89.48
$(L, V) \rightarrow L$	0.80	1.00	$(V, L) \rightarrow V$	0.80	1.00	$(A, L) \rightarrow A$	0.80	1.00	89.62	89.60
$(L, A) \rightarrow L$	0.00	0.80	$(V, A) \rightarrow V$	0.00	0.80	$(A, V) \rightarrow A$	0.00	0.80	88.87	88.85
$(L, A) \rightarrow L$	0.00	0.50	$(V, A) \rightarrow V$	0.00	0.50	$(A, V) \rightarrow A$	0.00	0.50	89.95	89.92
$(L, A) \rightarrow L$	0.00	0.20	$(V, A) \rightarrow V$	0.00	0.20	$(A, V) \rightarrow A$	0.00	0.20	88.56	88.55
$(V, A) \rightarrow L$	0.20	0.20	$(L, A) \rightarrow V$	0.20	0.20	$(L, V) \rightarrow A$	0.20	0.20	89.01	88.98
$(V, A) \rightarrow L$	0.50	0.50	$(L, A) \rightarrow V$	0.50	0.50	$(L, V) \rightarrow A$	0.50	0.50	88.85	88.83
$(V, A) \rightarrow L$	0.80	0.80	$(L, A) \rightarrow V$	0.80	0.80	$(L, V) \rightarrow A$	0.80	0.80	89.31	89.26
$(L, V, A) \rightarrow L$	0.20	0.80	$(L, V, A) \rightarrow V$	0.20	0.80	$(L, A, V) \rightarrow A$	0.20	0.80	90.13	90.04
$(L, V, A) \rightarrow L$	0.30	0.70	$(L, V, A) \rightarrow V$	0.30	0.70	$(L, A, V) \rightarrow A$	0.30	0.70	**90.41**	**90.38**
$(L, V, A) \rightarrow L$	0.40	0.60	$(L, V, A) \rightarrow V$	0.40	0.60	$(L, A, V) \rightarrow A$	0.40	0.60	90.20	90.17

the best performance with the BERT and COCOLM, which further verify the effectiveness of our proposed TMMDA.

4.3.2 Token Mixup variants. To demonstrate the effectiveness of our proposed CTM, we introduce multiple variants to support the design choice of CTM, which employs uni-modality, bi-modality, and tri-modality to perform data augmentation, respectively. According to Equations 4, 6 and 8, we reduce the token ratio of one or more modalities by controlling the mixing ratio ρ_{la}, ρ_{lv}, ρ_{vl}, ρ_{va}, ρ_{al}, and ρ_{av}. The arrow '→' indicates the direction of information flow that enhances the raw modality with the virtual modality. For example, '$(L, V, A) \longrightarrow L$' means using text, visual, acoustic modalities to generate virtual modality and enhance text modality. From the Table. 4, we observed that TMMDA achieves the best performance under the condition of the tri-modality mixup. Uni-modality mixup does not perform as well as the bi-modality. These results demonstrate that TMMDA can explore different data augmentation patterns by controlling the mixup ratio, enabling the model to improve its generalization ability of the model.

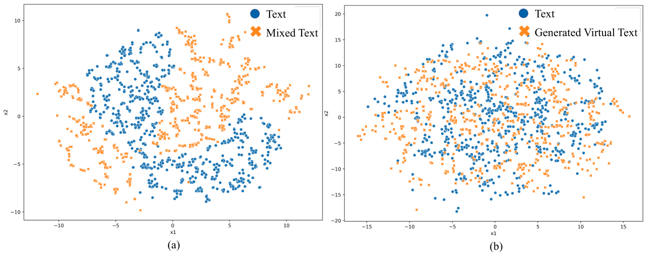

Figure 4: t-SNE visualization of mixed and generated virtual text representations on CMU-MOSI.

4.3.3 Module Analysis. To gain insights into our four parts, we conducted ablation studies incrementally. Concretely, we compared our model TMMDA with the following variants: 1) w/o CTM, we remove the cross-modal token mixup, and only employ random noise as input of the generator; 2) w/o CGAN, we eliminate the cross-modal generative adversarial network, and only take mixed representation as the input of the cross-modal encoder module. 3) w/o CME, without the cross-modal encoder and concatenating the representations of raw modality and generated virtual modality in the token dimension; and 4) w/o JSD, excluding the Jensen–Shannon divergence. As shown in Table 3, compared with the TMMDA, the absence of the CGAN module results in sharp performance degradation. Specifically, it drops absolutely by 1.86% and 1.84% on Acc2 and F1-Score on CMU-MOSI for MSA respectively. This demonstrates the vital importance of generated virtual modalities as it can learn informative multimodal representations. Besides, TMMDA achieves better performance than w/o CTM, revealing that the cross-modal token mixup can help the generator to learn modality-relevant information and improve model performance. Moreover, the results drop of w/o CME can be observed, indicating that it is important to consider supplementing the raw modality information with generated virtual modality information. Generally, our proposed model TMMDA achieves the best performance compared with all variants on CMU-MOSI, verifying the effectiveness and complementarity of four parts.

4.4 Qualitative Analysis

To qualitatively validate the effectiveness of TMMDA, we showed several typical examples of sentiment prediction (i.e., positive, negative, neutral sentiments) in Table 5. Based on these sentiment analysis results, we could draw a conclusion that our model could comprehend positive, negative and neutral sentiments accurately. In case A, case B, and case C, the text, visual, acoustic modalities are seen to provide essential information for multimodal sentiment analysis. Our model is able to predict values that are very close to the ground truth. Visual and acoustic modalities in Case D do not

Table 5: Input and predictions of four samples in our case study on CMU-MOSI dataset.

Case	Text	Visual	Acoustic	Prediction	Truth
A	The verdict is stupid and a complete waste of money.	Open Wide	Pause	-2.63	-2.59 ✓
B	The um cross of personality is really um charismatic and dynamic.	Relaxed look	Rhythm changes	+1.95	+2.00 ✓
C	Or big collector of the action figures.	No expression	Normal Voice	-0.01	+0.0 ✓
D	Even tell funny jokes.	Reply disdainfully	Particular tone	+1.47	-1.79 ×

provide obvious discriminative information for sentiment analysis, and the bias of the pre-trained language representation model misleads the prediction results of the model. The result demonstrates that it is still challenging to overcome the language bias problem on existing datasets.

To explore the impact of the generative model, we used t-SNE [29] to visualize the distribution of mixed and the generated virtual modality representation. The results are shown in Fig. 4. From the comparison results, we could find that there is a minor distributional modality gap between the mixed sequence and the raw modality i.e., sub-figure (a). This result shows that cross-modal token mixup can retain the semantic information of the raw modalities. We employ the CGAN and JSD to further reduce the gap between generated virtual modality and raw modality for fusion, i.e., sub-figure (b), which helps the cross-modal encoder to enhance raw modality via generated virtual modality.

5 CONCLUSION

In this paper, we introduced TMMDA, a new token mixup multimodal data augmentation for multimodal sentiment analysis tasks, which consists of cross-modal token mixup, cross-modal generative adversarial network, and cross-modal encoder modules. These modules cooperate to create new training samples and generate new virtual modalities, and then enhance raw modalities. Experimental results and analysis on both CMU-MOSI and CMU-MOSEI datasets verify the effectiveness and generalization of our proposed method.

ACKNOWLEDGMENTS

This study is partially supported by National Key R&D Program of China (2021ZD0113402), National Natural Science Foundations of China (62276082, U1813215 and 61876052), National Natural Science Foundation of Guangdong, China (2019A1515011158), Major Key Project of PCL (PCL2021A06), Strategic Emerging Industry Development Special Fund of Shenzhen (20200821174109001) and Pilot Project in 5G + Health Application of Ministry of Industry and Information Technology & National Health Commission (5G + Luohu Hospital Group: an Attempt to New Health Management Styles of Residents). This work is also supported by Education Center of Experiments and Innovations at Harbin Institute of Technology, Shenzhen.

REFERENCES

[1] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5799–5809.

[2] Alex Falcon, Giuseppe Serra, and Oswald Lanz. 2022. A Feature-space Multimodal Data Augmentation Technique for Text-video Retrieval. *arXiv preprint arXiv:2208.02080* (2022).

[3] Qingkai Fang and Yang Feng. 2022. Neural Machine Translation with Phrase-Level Universal Visual Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5687–5698.

[4] Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7050–7062.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[6] Dengji Guo, Zhengrui Ma, Min Zhang, and Yang Feng. 2022. Prediction Difference Regularization against Perturbation for Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7665–7675.

[7] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis Philippe Morency, and Soujanya Poria. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *ICMI 2021-Proceedings of the 2021 International Conference on Multimodal Interaction*. Association for Computing Machinery, Inc, 6–15.

[8] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9180–9192.

[9] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2022. MixGen: A New Multi-Modal Data Augmentation. *arXiv preprint arXiv:2206.08358* (2022).

[10] 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*. 1122–1131.

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[12] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3405–3424.

[13] Caterina Lacerra, Rocco Tripodi, and Roberto Navigli. 2021. GENESIS: A Generative Approach to Substitutes in Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10810–10823.

[14] Jicheng Li, Pengzhi Gao, Xuanfu Wu, Yang Feng, Zhongjun He, Hua Wu, and Haifeng Wang. 2021. Mixup Decoding for Diverse Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 312–320.

[15] Chengliang Liu, Zhihao Wu, Jie Wen, Yong Xu, and Chao Huang. 2022. Localized sparse incomplete multi-view clustering. *IEEE Transactions on Multimedia* (2022).

[16] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 8 (2020), 726–742.

[17] 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)*. 2247–2256.

[18] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2554–2562.

[19] Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu. 2021. MixSpeech: Data augmentation for low-resource automatic speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7008–7012.

[20] Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems* 34 (2021), 23102–23114.

[21] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*. PMLR, 2642–2651.

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[23] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359.

[24] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems* 34 (2021), 13937–13949.

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

[26] Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N Metaxas. 2020. Selfnorm and crossnorm for out-of-distribution robustness. (2020).

[27] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[28] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In *International Conference on Representation Learning*.

[29] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

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

[31] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*. PMLR, 6438–6447.

[32] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. *Advances in neural information processing systems* 29 (2016).

[33] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. 2021. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12206–12215.

[34] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. 2022. Multimodal Token Fusion for Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12186–12195.

[35] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223.

[36] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *arXiv e-prints* (2019), arXiv–1901.

[37] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33 (2020), 6256–6268.

[38] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1009–1021.

[39] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3995–4007.

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

[41] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.

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

[43] 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).

[44] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.

[45] Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is Making the Contribution: Modulating Unimodal and Cross-modal Dynamics for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 1262–1274.

[46] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[47] Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2021. A Language Model-based Generative Classifier for Sentence-level Discourse Parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2432–2446.

[48] Xianbing Zhao, Yixin Chen, Wanting Li, Lei Gao, and Buzhou Tang. 2022. MAG+: An Extended Multimodal Adaptation Gate for Multimodal Sentiment Analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4753–4757.

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13001–13008.

A HYPERPARAMETER SETTING

Table. A1 shows the performance difference between TMMDA and the well-known models MAG and MulT on the low-resource data scenario. Concretely, we randomly sample 100%, 70%, and 50% ratio of the training samples to participate in the training of the model. By observing Table. A1, we conclude that our model still performs better than MAG and MulT in the scenario of low resource data. Table. A2 displays the experimental settings of our proposed TMMDA that we train on multimodal sentiment analysis tasks.

Table A1: Ablation experiments of TMMDA on the CMU-MOSI dataset. The best results are highlighted in bold.

Model	Ratio	Acc-2	F1-Score	MAE	CC
TMMDA	100%	90.41	90.38	0.593	0.870
TMMDA	70%	89.62	89.60	0.628	0.861
TMMDA	50%	89.44	89.43	0.640	0.855
MAG	100%	88.70	88.53	0.624	0.857
MAG	70%	87.79	87.78	0.684	0.842
MAG	50%	87.33	87.31	0.686	0.839
MuLT	100%	88.55	88.52	0.654	0.856
MuLT	70%	87.34	87.32	0.683	0.847
MuLT	50%	87.03	87.01	0.698	0.843

Table A2: The hyperparameter settings used in CMU-MOSI and CMU-MOSEI benchmark.

Hyperparameter	CMU-MOSI	CMU-MOSEI
Batch Size/Epochs	128/150	64/150
Optimizer	Adam	Adam
Learning Rate	6e-5	1e-4
$\beta_{dist}/\beta_{gan}/\beta_{task}$	0.5/0.5/1	0.5/0.5/1
$\alpha_l/\alpha_v/\alpha_a$	0.1/0.1/0.1	0.1/0.1/0.1