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

https://github.com/xiaobaicaihhh/TMMDA.

—— WWW 2023

2023.5.11 • ChongQing

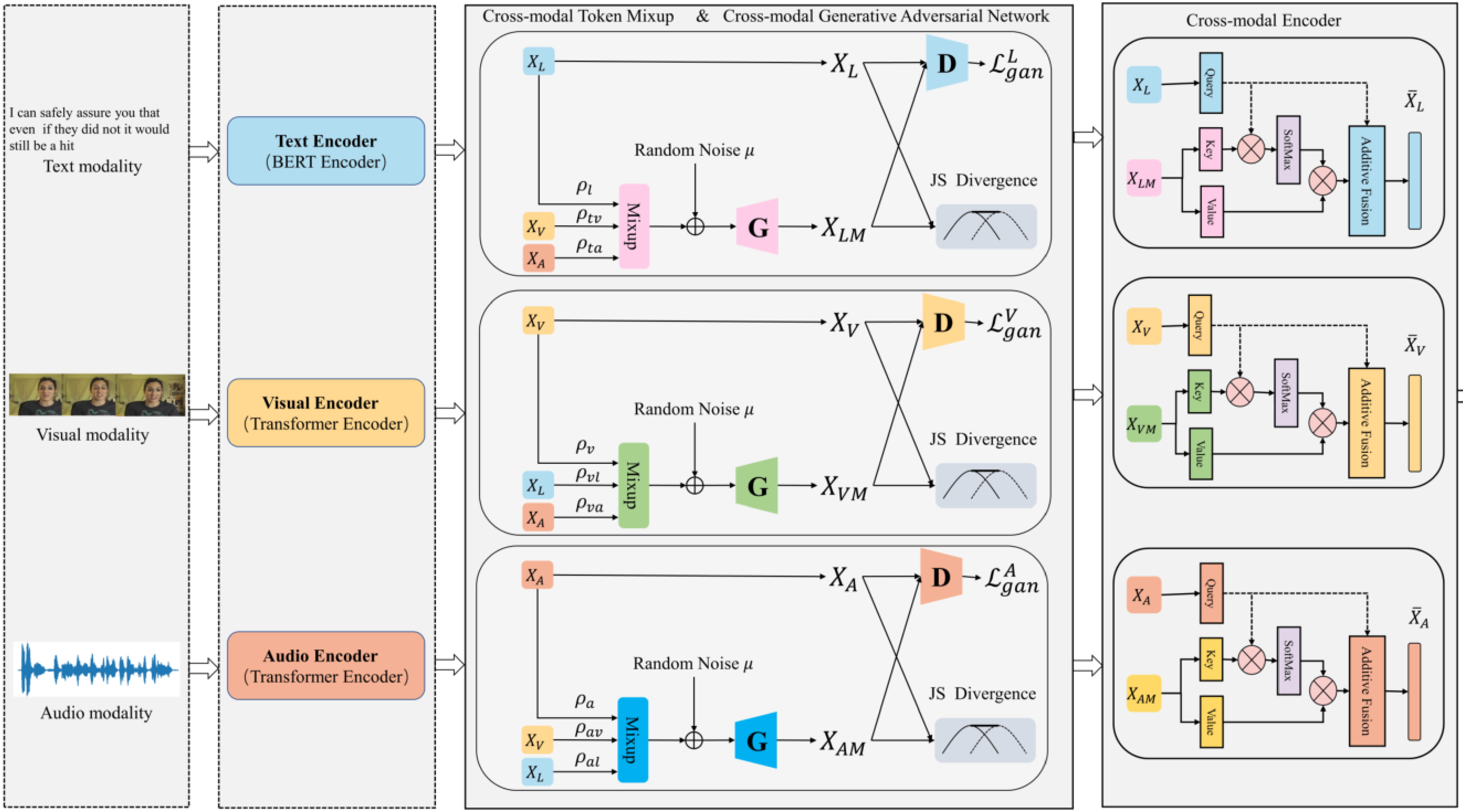
Leibniz-Institut für Sozialwissenschaften

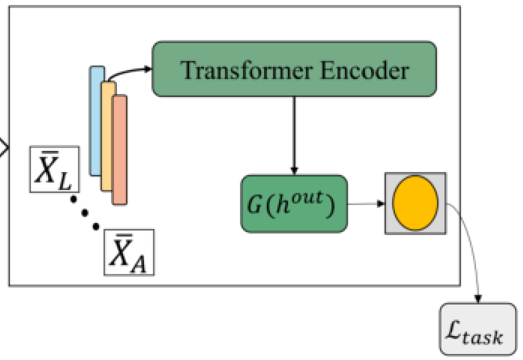
Reported by JiaWei Cheng

Motivation

Data augmentation training strategy has achieved great success to improve model performance in multiple computer vision (CV) and natural language processing (NLP) tasks. However, it is not straightforward to apply previous data augmentation methods for multimodal sentiment analysis tasks

Overview





$$X_L = \{x_{l_1}, x_{l_2}, \ldots, x_{l_n}\} \quad (1)$$

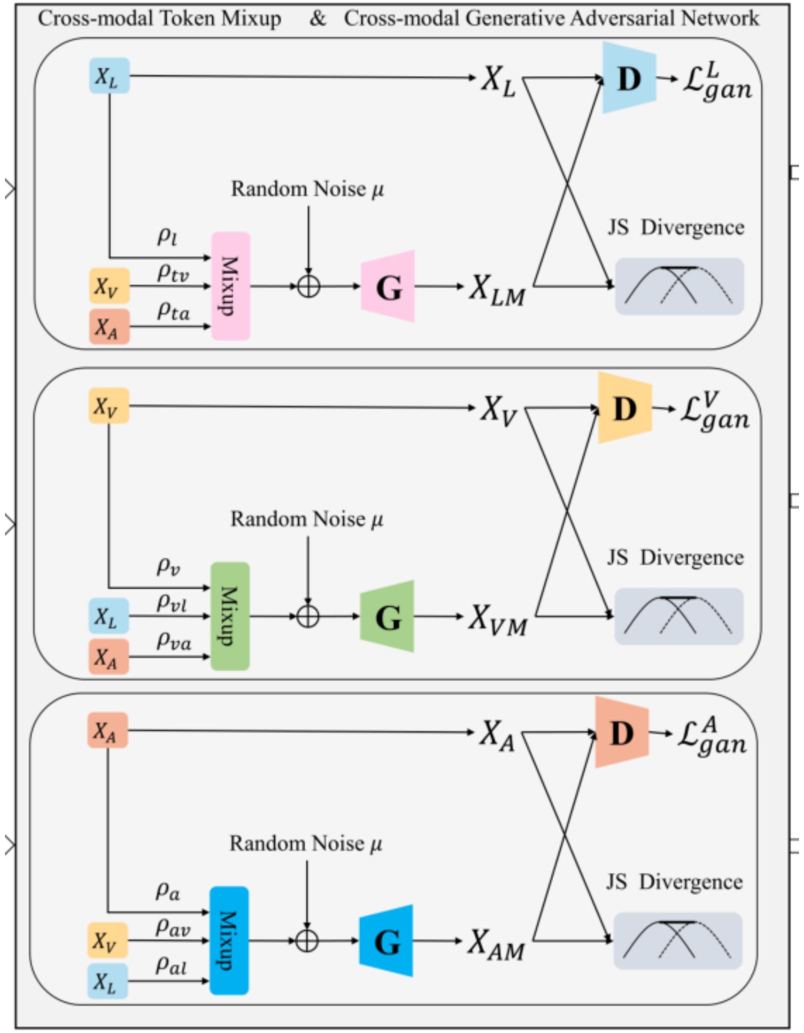
$$X_V = \{x_{v_1}, x_{v_2}, \ldots, x_{v_n}\} \quad (2)$$

$$X_A = \{x_{a_1}, x_{a_2}, \ldots, x_{a_n}\} \quad (3)$$

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

$$\dot{X}_{LM} = \{x_{lm_1}, x_{lm_2}, \ldots, x_{lm_n}\} \quad (5)$$

Method



$$X_L = \{x_{l_1}, x_{l_2}, \ldots, x_{l_n}\} \quad (1)$$

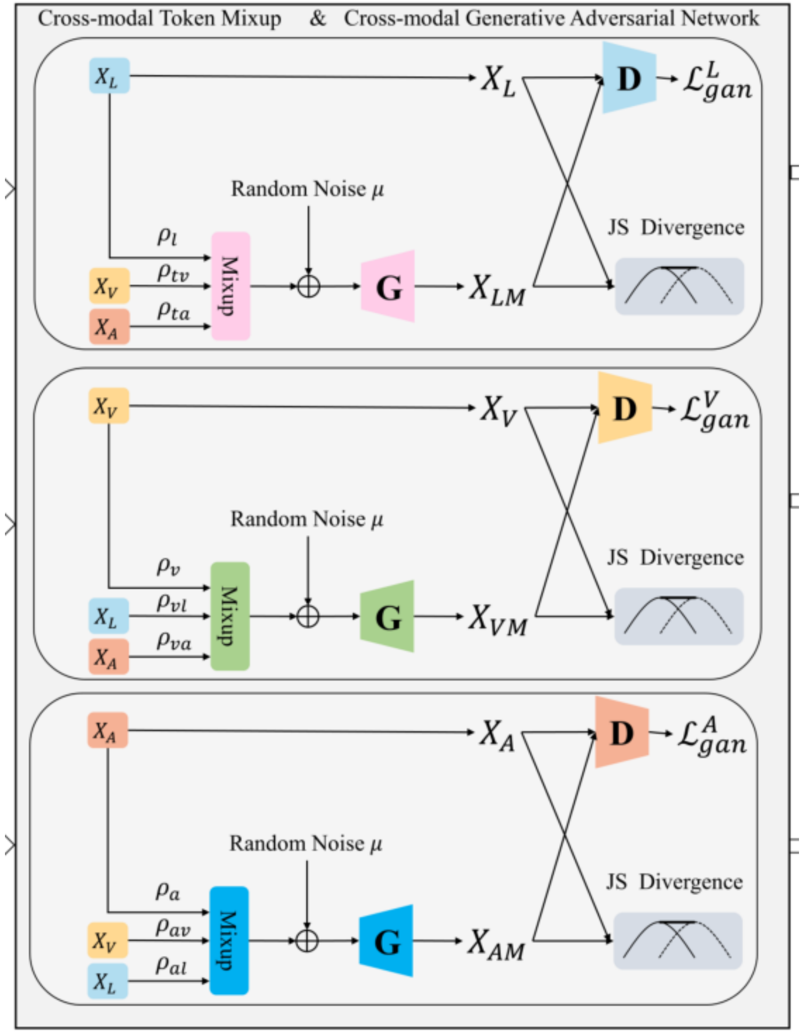
$$X_V = \{x_{v_1}, x_{v_2}, \ldots, x_{v_n}\} \quad (2)$$

$$X_A = \{x_{a_1}, x_{a_2}, \ldots, x_{a_n}\} \quad (3)$$

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

$$\dot{X}_{LM} = \{x_{lm_1}, x_{lm_2}, \ldots, x_{lm_n}\} \quad (5)$$

Method



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

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

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

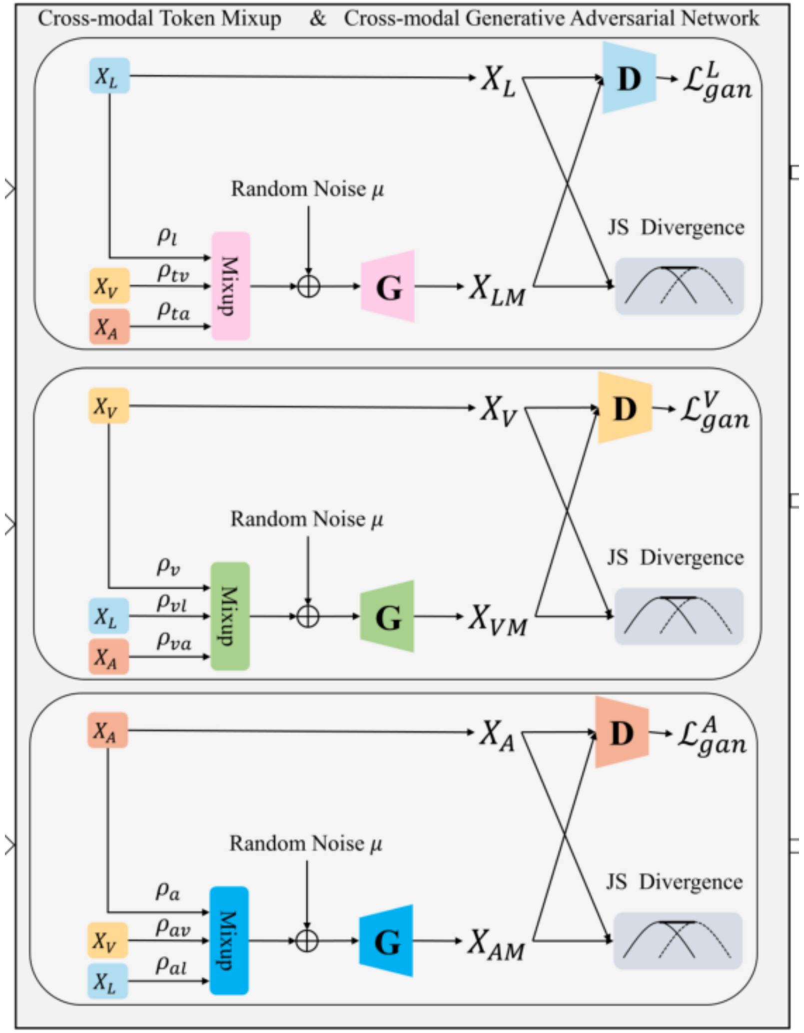
$$\dot{X}_{AM} = \{x_{am_1}, x_{am_2}, \dots, x_{am_n}\} \tag{9}$$

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

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

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

Method



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

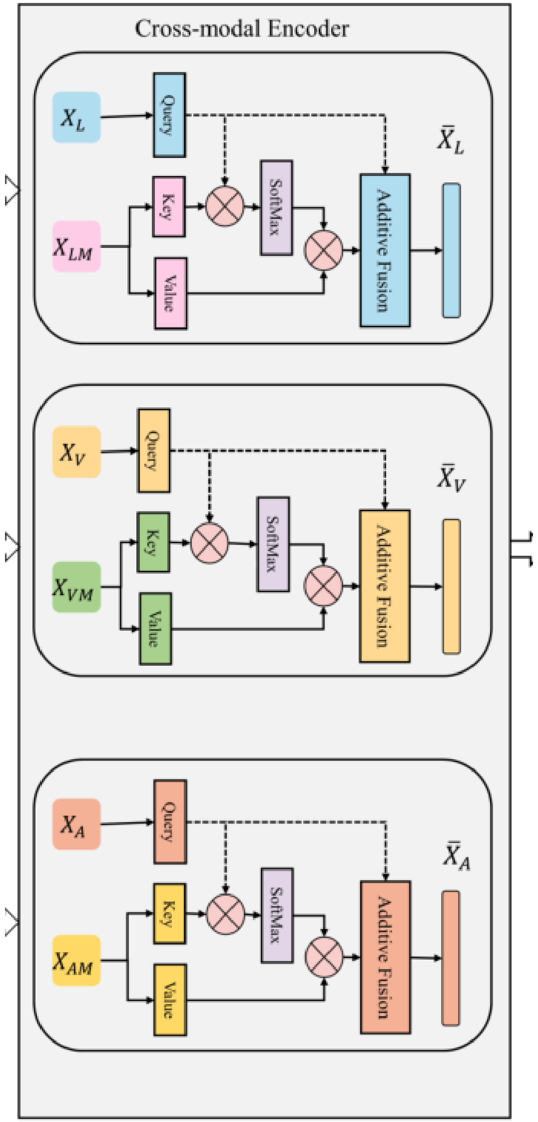
$$D_{KL}(p\|q) = -\sum_{x} p(x)\log\frac{q(x)}{p(x)} \tag{14}$$

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

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

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

Experiments

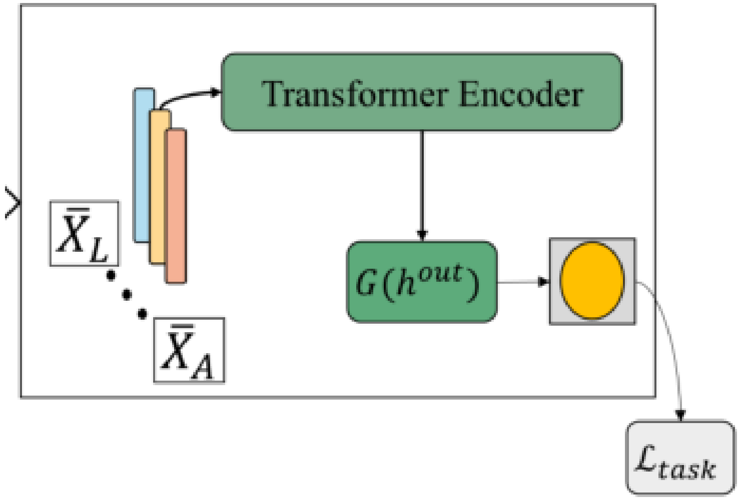


$$\begin{aligned} Y_{S1} &= \mathrm{CM}_{S2\rightarrow S1}\left(X_{S1}, X_{S2}\right) \\ &= \operatorname{softmax}\left(\frac{Q_{S1}K_{S2}^{\top}}{\sqrt{d_k}}\right)V_{S2} \\ &= \operatorname{softmax}\left(\frac{X_{S1}W_{Q_{S1}}W_{K_{S2}}^{\top}X_{S2}^{\top}}{\sqrt{d}}\right)X_{S2}W_{V_{S2}} \end{aligned}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

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

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

Experiments



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

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

Experiments

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**

Experiments

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**

Experiments

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**

Experiments

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}	ρ_{va}	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

Experiments

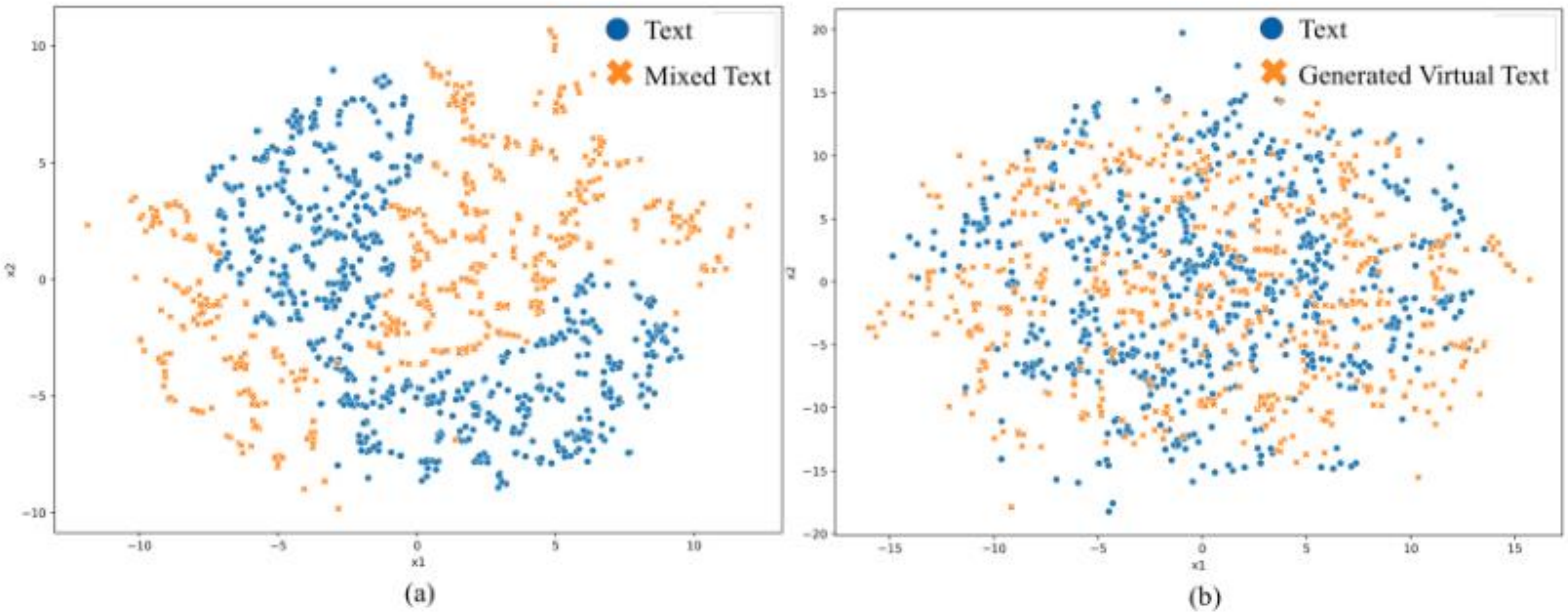


Figure 4: t-SNE visualization of mixed and generated virtual text representations on CMU-MOSI.

Experiments

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 ✗

Thanks!