Learning Robust Multi-Modal Representation for Multi-Label Emotion Recognition via Adversarial Masking and Perturbation

Shipping Ge
shipingge@smail.nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China

Zhiwei Jiang*
jzw@nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China

Zifeng Cheng
chengzf@smail.nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China

Cong Wang
cw@smail.nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China

Yafeng Yin
yafeng@nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China

Qing Gu
guq@nju.edu.cn
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China

https://github.com/ShipingGe/MMER

—— WWW 2023

Reported by Yuyang Lai

1.Introduction

2.Method

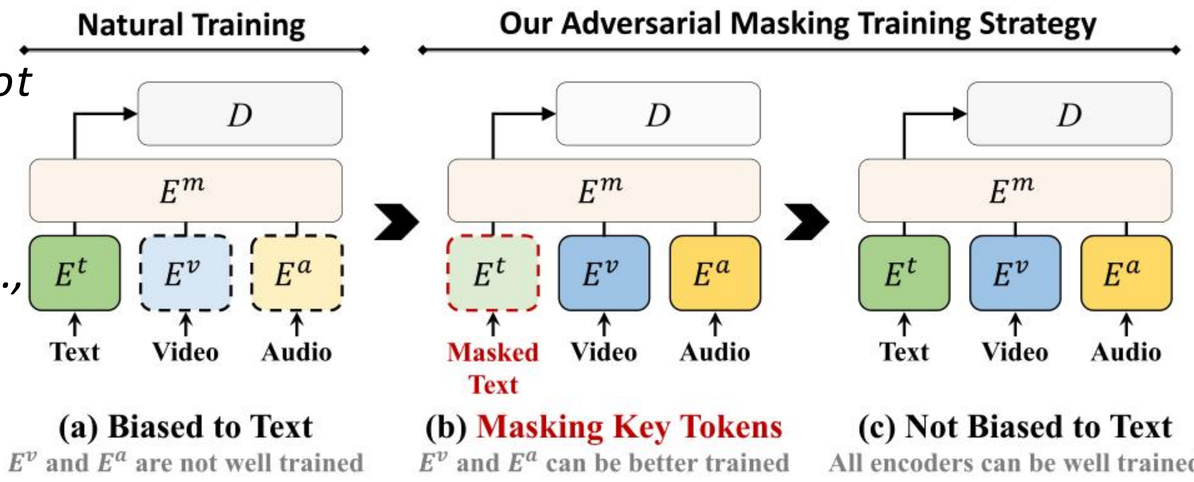
3.Experiments

Introduction

PROBLEMS:

modality bias of representation——*natural training does not guarantee that every modality can be adequately encoded.*

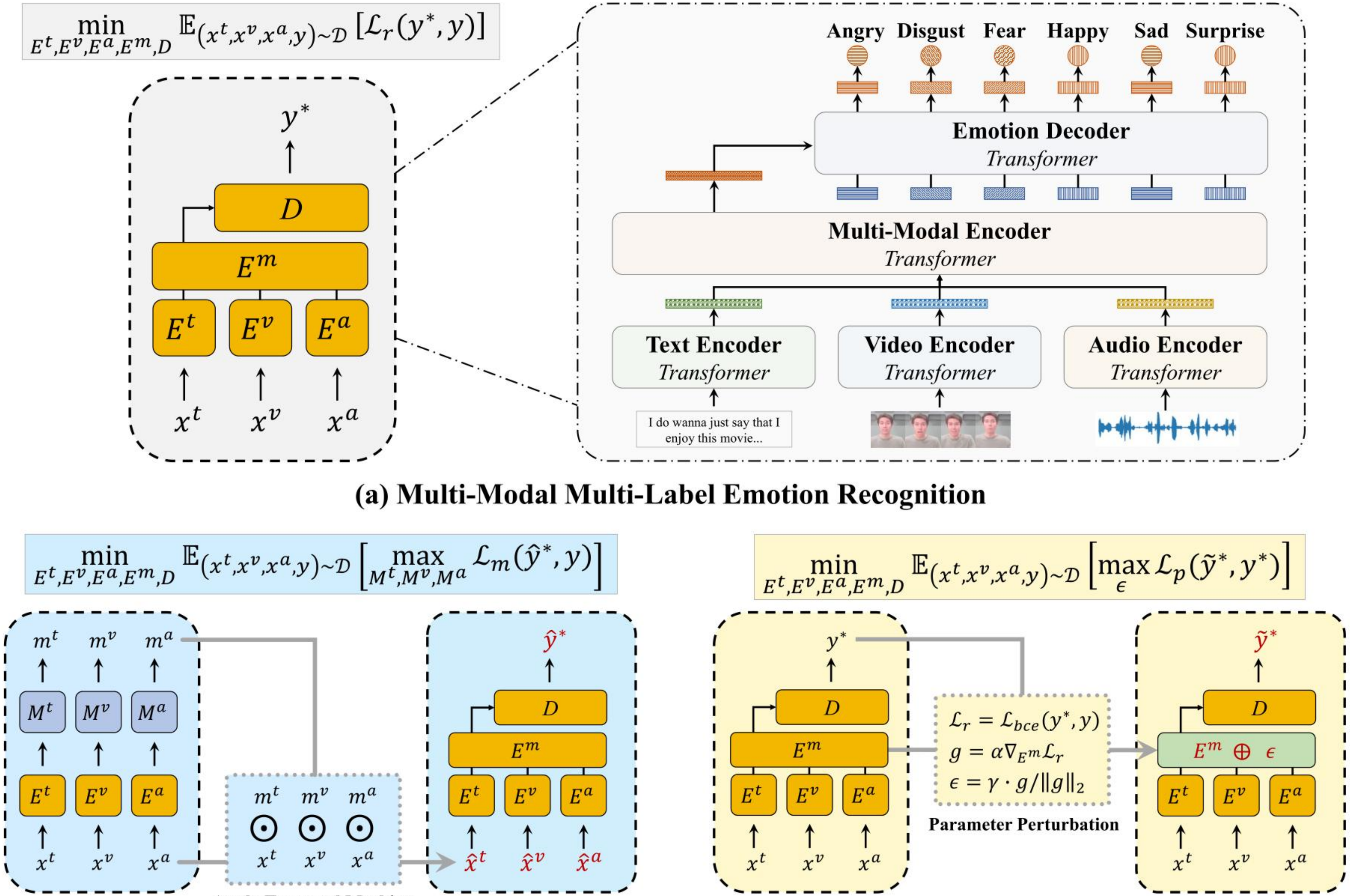
data bias of training——*natural training without intervention (e.g., regularization) may cause the model overfit the training data.*



SOLUTIONS:

temporal masking strategy, *aiming at enhancing the encoding of other modalities by masking the most emotion-related temporal units (e.g., words for text or frames for video) of the informative modality.*

parameter perturbation strategy,*aiming at enhancing the generalization of the model by adding the adversarial perturbation to the intermediate parameters of model as model regularization.*

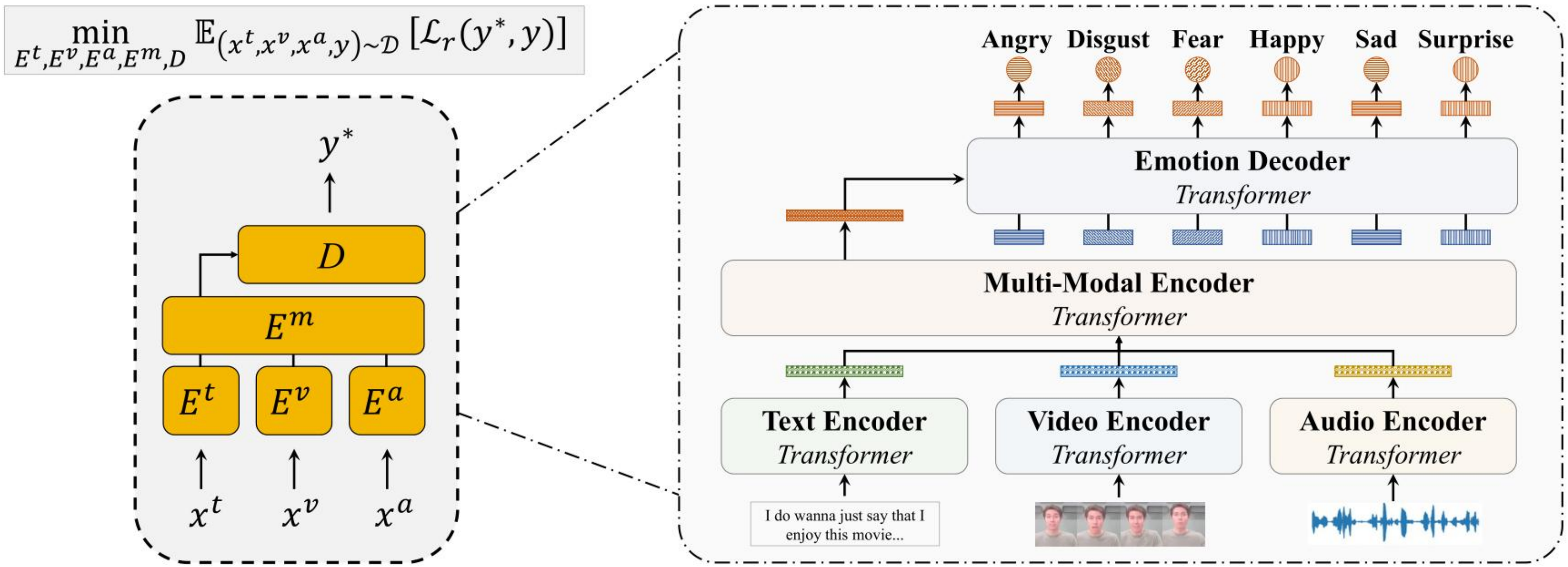


(a) Multi-Modal Multi-Label Emotion Recognition

(b) Adversarial Temporal Masking

(c) Adversarial Parameter Perturbation

Method



(a) Multi-Modal Multi-Label Emotion Recognition

$\mathcal{Y} = \{1, 2, \ldots, K\}$

$\mathcal{D} = \{(\boldsymbol{x}_i^t, \boldsymbol{x}_i^v, \boldsymbol{x}_i^a), \boldsymbol{y}_i\}_{i=1}^N$ $\quad$ $\boldsymbol{x}_i^t = \{\boldsymbol{x}_{ij}^t\}_{j=1}^{L^t}$ $\quad$ $\boldsymbol{x}_i^a = \{\boldsymbol{x}_{ij}^a\}_{j=1}^{L^a}$ $\quad$ $\boldsymbol{x}_i^v = \{\boldsymbol{x}_{ij}^v\}_{j=1}^{L^v}$

$E = \{E^t, E^v, E^a, E^m\}$ $\quad$ $\theta = \{\theta_E, \theta_D\}$

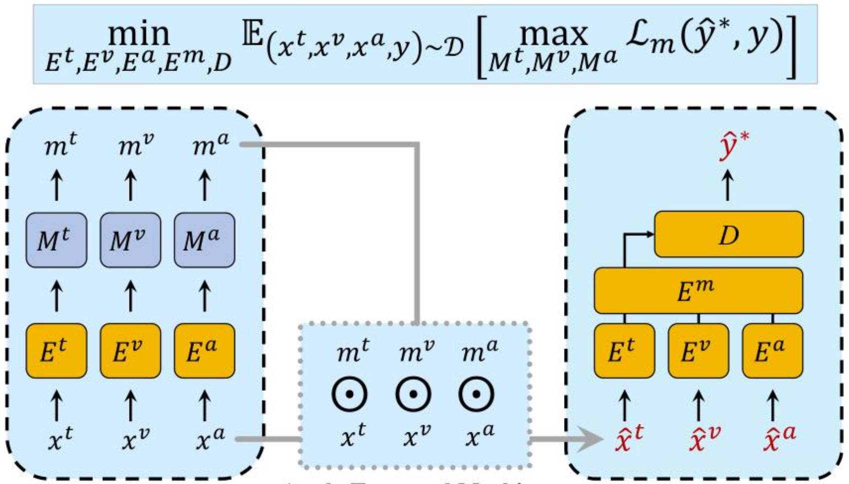
$\boldsymbol{H}^c = [\boldsymbol{H}^t, \boldsymbol{H}^v, \boldsymbol{H}^a]$

$$\min_{E,D} \mathbb{E}_{(\boldsymbol{x}^t,\boldsymbol{x}^v,\boldsymbol{x}^a,\boldsymbol{y})\sim\mathcal{D}}[\mathcal{L}_r(\boldsymbol{y}^*, \boldsymbol{y})] \tag{1}$$

$$\mathcal{L}_r(\boldsymbol{y}^*, \boldsymbol{y}) = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathcal{L}_{bce}(\boldsymbol{y}_{ij}^*, \boldsymbol{y}_{ij}) \tag{2}$$

$$\mathcal{L}_{bce}(y^*, y) = y \log y^* + (1 - y) \log(1 - y^*) \tag{3}$$

Method



(b) Adversarial Temporal Masking

$$\min_{E,D} \mathbb{E}_{(\boldsymbol{x}^t,\boldsymbol{x}^v,\boldsymbol{x}^a,\boldsymbol{y})\sim\mathcal{D}}\left[\max_{M^t,M^v,M^a} \mathcal{L}_m(\hat{\boldsymbol{y}}^*, \boldsymbol{y})\right] \tag{4}$$

$$\mathcal{L}_m(\hat{\boldsymbol{y}}^*, \boldsymbol{y}) = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathcal{L}_{bce}(\hat{\boldsymbol{y}}^*_{ij}, \boldsymbol{y}_{ij}) \tag{5}$$

$$\boldsymbol{m} = \mathcal{T}(\mathrm{Sigmoid}(\boldsymbol{HW} + \boldsymbol{b})) \tag{6}$$

$$\mathcal{T}(\boldsymbol{s})_i = \begin{cases} 0, & \boldsymbol{s}_i \in k \text{ highest candidates.} \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

$$\Gamma^* = \arg\max_{\Gamma \geq 0} < C, \Gamma > \quad \text{s.t.} \quad \Gamma \mathbf{1}_m = \mathbf{1}_n / n, \Gamma^\top \mathbf{1}_n = [k/n, (n-k)/n] \tag{8}$$

$$\mathcal{T} = n\Gamma^* \cdot [1, 0]^\top \tag{9}$$

$$\hat{\boldsymbol{x}} = \boldsymbol{x} \odot \boldsymbol{m} \tag{10}$$

$$\theta_M \leftarrow \theta_M + \alpha \nabla_{\theta_M} \mathcal{L}_m(\hat{\boldsymbol{y}}^*, \boldsymbol{y}) \tag{11}$$

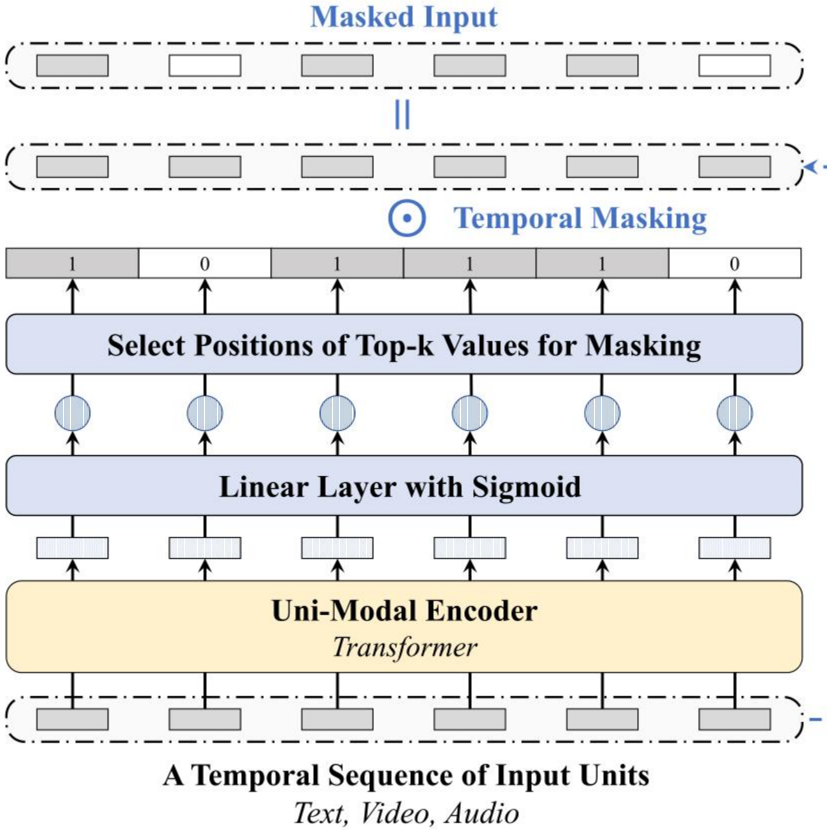
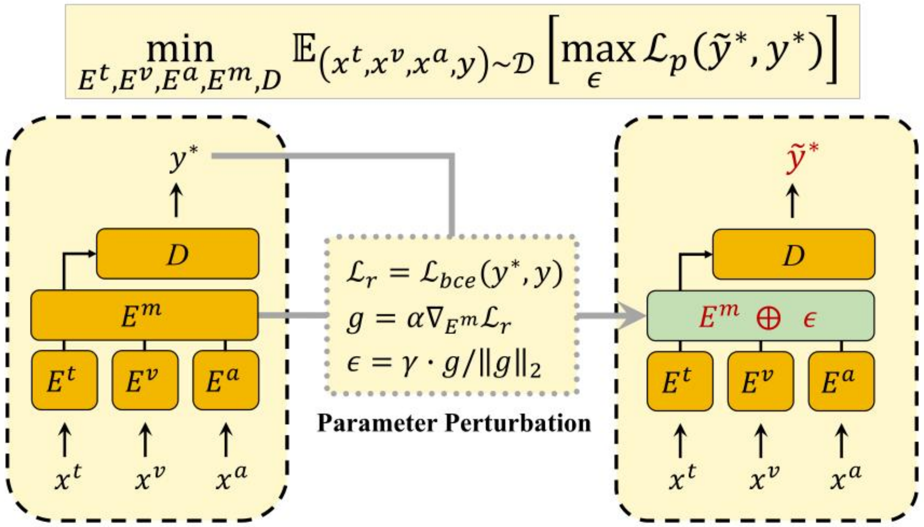


Figure 3: The detailed illustration of ATM strategy.

Method



(c) Adversarial Parameter Perturbation

$$\min_{E,D} \mathbb{E}_{(\boldsymbol{x}^t,\boldsymbol{x}^v,\boldsymbol{x}^a,\boldsymbol{y})\sim\mathcal{D}}\Big[\max_{\epsilon,\|\epsilon\|\le\gamma} \mathcal{L}_p(\tilde{\boldsymbol{y}}^*, \boldsymbol{y}^*)\Big] \tag{12}$$

$$\begin{aligned}\mathcal{L}_p(\tilde{\boldsymbol{y}}^*, \boldsymbol{y}^*) = \mathrm{KL}(p(\boldsymbol{y}^*|\boldsymbol{x}^t, \boldsymbol{x}^v, \boldsymbol{x}^a; \theta)\| \\ p(\tilde{\boldsymbol{y}}^*|\boldsymbol{x}^t, \boldsymbol{x}^v, \boldsymbol{x}^a; \theta+\epsilon))\end{aligned} \tag{13}$$

$$\epsilon = \gamma \cdot \frac{g}{\|g\|_2}, \text{ where } g = \alpha\nabla_{\theta_{E^m}}\mathcal{L}_r \tag{14}$$

$$\begin{aligned}\min_{E,D} \mathbb{E}_{(\boldsymbol{x}^t,\boldsymbol{x}^v,\boldsymbol{x}^a,\boldsymbol{y})\sim\mathcal{D}}\Big[\mathcal{L}_r(\boldsymbol{y}^*, \boldsymbol{y}) \\ + \rho \max_{M} \mathcal{L}_m(\hat{\boldsymbol{y}}^*, \boldsymbol{y}^*) + \sigma \max_{\epsilon,\|\epsilon\|\le\gamma} \mathcal{L}_p(\tilde{\boldsymbol{y}}^*, \boldsymbol{y}^*)\Big]\end{aligned} \tag{15}$$

Experiments

Approach	Methods	CMU-MOSEI				NEMu			
		$Acc(\%)$	HL	$miF_1(\%)$	$maF_1(\%)$	$Acc(\%)$	HL	$miF_1(\%)$	$maF_1(\%)$
Classical	**BR** (Shen et al. 2003)	22.2	0.371	38.6	34.7	23.0	0.475	41.1	40.5
	CC (Read et al. 2011)	22.5	0.377	38.6	34.1	23.5	0.465	41.7	41.1
	LP (Tsoumakas et al. 2010)	15.9	0.426	28.6	28.8	21.1	0.414	37.2	35.0
Linguistic	**LASN** (Xiao et al. 2019)	39.3	0.209	50.1	32.3	19.5	0.332	39.7	35.7
	Seq2set (Yang et al. 2019)	45.7	0.231	53.8	34.0	24.8	0.424	42.1	39.7
	KRF (Ma et al. 2020)	45.3	0.226	51.5	29.0	23.1	0.496	42.0	39.7
Non-linguistic	**ML-GCN** (Chen et al. 2019)	41.1	0.207	50.9	29.7	15.8	0.293	34.4	27.8
	MLEE (Ando et al. 2019)	43.7	0.211	52.8	38.6	-	-	-	-
Muti-modal	**MulT** (Tsai et al. 2019)	44.5	0.190	53.1	34.4	17.9	0.293	42.6	39.0
	CIA (Chauhan et al. 2019)	42.9	0.214	45.5	11.7	11.1	0.336	29.6	34.0
	M3ER (Mittal et al. 2020)	40.9	0.195	51.9	34.9	19.4	0.281	40.6	36.4
	HHMPN (Zhang et al. 2021)	45.9	0.189	55.6	**43.0**	24.9	**0.270**	46.1	43.5
	TAILOR† (Zhang et al. 2022)	43.7	0.206	49.7	37.1	21.6	0.281	40.6	35.9
	Ours	**48.4**	**0.185**	**56.9**	41.7	**30.3**	0.291	**50.2**	**47.4**

Table 1: Comparison of our method with the existing emotion recognition methods on the CMU-MOSEI dataset and NEMu dataset. The best results are marked in bold. †: Since the threshold of prediction in the TAILOR method is 0.35, which is different from the commonly used 0.5 in other multi-label learning methods, we change their threshold to 0.5 and rerun their code for a fair comparison.

Experiments

Model Setting	CMU-MOSEI				NEMu			
	Acc(%)	HL	miF_1(%)	maF_1(%)	Acc(%)	HL	miF_1(%)	maF_1(%)
Full Model	**48.4**	**0.185**	**56.9**	**41.7**	**30.3**	**0.291**	**50.2**	**47.4**
$-E^t$ and x^t	44.1	0.224	51.2	34.1	23.6	0.336	44.6	42.1
$-E^v$ and x^v	46.6	0.217	53.4	36.5	29.1	0.304	48.9	46.3
$-E^a$ and x^a	47.3	0.199	54.1	38.7	26.1	0.319	46.0	43.1
$-E^m$	47.4	0.203	53.5	38.4	28.1	0.301	48.7	45.5
$-D$(+ classifier over E^m)	46.9	0.190	54.1	39.3	29.1	0.299	48.9	46.7
−ATM strategy	46.2	0.203	52.8	36.1	27.6	0.314	48.1	46.5
$-M^t, M^v, M^a$(+ random mask)	47.1	0.197	53.5	36.4	27.3	0.316	47.9	45.5
−adversarial gradient reversal	47.6	0.196	54.1	36.6	28.5	0.299	48.6	46.7
−APP strategy	46.9	0.196	53.4	37.9	28.8	0.293	48.9	46.1
$-g$ (+random perturbation)	46.7	0.201	52.9	36.1	26.1	0.317	45.7	41.3
$-Attention_\epsilon$	47.5	0.189	54.3	38.4	29.1	0.291	49.6	46.4
$-\mathrm{FNN}_\epsilon$	48.0	0.185	55.0	39.8	30.0	0.303	49.3	47.0
−Transformer (+LSTM)	45.2	0.217	53.6	36.4	25.4	0.314	45.1	42.2
−ATM, APP, Transformer (+LSTM)	41.7	0.238	49.1	34.5	22.1	0.323	43.4	41.1
−ATM and APP strategies	45.5	0.210	52.1	35.5	25.4	0.296	47.3	44.9
−ATM and APP strategies (+ FGSM)	46.2	0.207	53.8	35.1	26.5	0.299	48.4	45.3
−ATM and APP strategies (+ PGD)	45.3	0.216	52.3	35.7	25.1	0.301	46.9	44.6
−ATM and APP (+ rand erasing)	44.2	0.221	50.1	34.1	25.5	0.311	47.1	45.1
−ATM and APP (+ Gaussian noise)	42.3	0.232	48.9	33.4	23.1	0.326	45.8	42.1

Table 2: Ablation study of our model. Accuracy, Precision, Recall, and Micro-F1 scores on the CMU-MOSEI dataset and NEMu dataset. '$-E^t$ and x^t' means removing the text encoder and text input. Note that for NEMu, $-E^t$ means removing lyrics and comments features at the same time.

Experiments

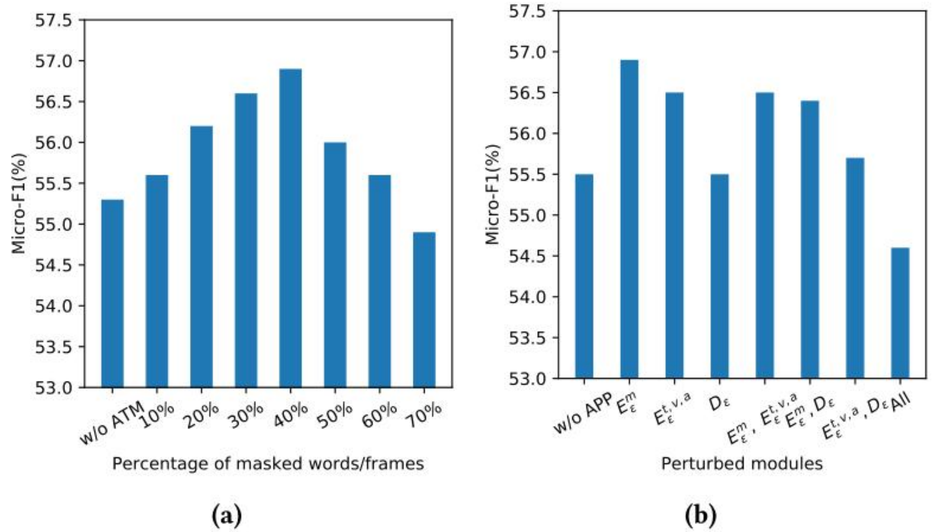


Figure 4: (a) Comparison of different percentages of masked units in ATM during training. (b) Comparison of applying APP to different modules of our model. $E_{\epsilon}^{t,v,a}$ denotes applying APP to the uni-modal encoders E^t, E^v, and E^a.

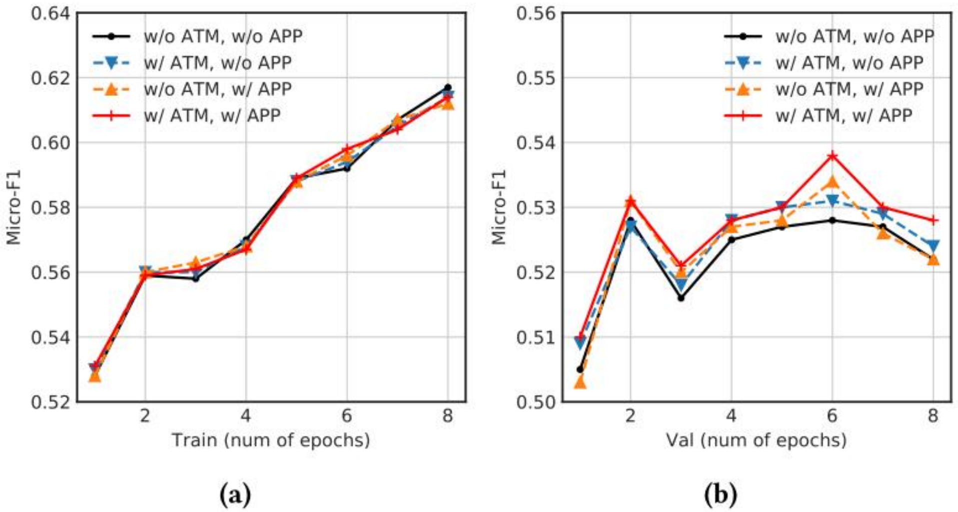


Figure 5: Comparison of different training strategies' performance on the training set and validation set of CMU-MOSEI at different training epochs.

Experiments

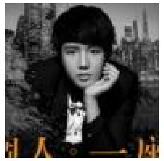
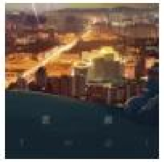
Images	*Lyrics*	*Comments*	Ground-Truth	Ours	HHMPN
	*Imagine if you can meet again at this moment, will you forget the past...*	*I never thought that so many friends have heard this song, and it is very touching...*	*Sad Lonely Miss Quiet Healing*	*Sad Lonely Miss Healing*	*Sad Lonely Healing*
	*Flashing the message of love, a few words into my heart, it is not easy to reveal the mood...*	*I've been sighing with emotion for such a beautiful song, I'm ecstatic, I'm unbelievably in love...*	*Happy Miss Romantic Refreshing*	*Happy Miss Healing Romantic Refreshing*	*Excited Happy Nostalgic Refreshing*
	*When I opened my eyes and went back to that year, I also had a crush on me...*	*I think back to the night he kissed me when I was eighteen...*	*Sad Lonely Quiet Miss Refreshing*	*Sad Lonely Quiet Miss Refreshing*	*Sad Lonely Quiet Moving Healing*

Figure 6: Cases of recognition results using our method and HHMPN. Audio is omitted here for simplicity.

Thank you!