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

Shiping 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

ABSTRACT

Recognizing emotions from multi-modal data is an emotion recognition task that requires strong multi-modal representation ability. The general approach to this task is to naturally train the representation model on training data without intervention. However, such natural training scheme is prone to modality bias of representation (i.e., tending to over-encode some informative modalities while neglecting other modalities) and data bias of training (i.e., tending to overfit training data). These biases may lead to instability (e.g., performing poorly when the neglected modality is dominant for recognition) and weak generalization (e.g., performing poorly when unseen data is inconsistent with overfitted data) of the model on unseen data. To address these problems, this paper presents two adversarial training strategies to learn more robust multi-modal representation for multi-label emotion recognition. Firstly, we propose an adversarial temporal masking strategy, which can enhance 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. Secondly, we propose an adversarial parameter perturbation strategy, which can enhance the generalization of the model by adding the adversarial perturbation to the parameters of model. Both strategies boost model performance on the benchmark MMER datasets CMU-MOSEI and NEMu. Experimental results demonstrate the effectiveness of the proposed method compared with the previous state-of-the-art method. Code will be released at https://github.com/ShipingGe/MMER.

CCS CONCEPTS

• **Information systems → Multimedia information systems**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

Emotion Recognition, Multi-Modal Learning, Multi-Label Learning, Adversarial Training

ACM Reference Format:
Shiping Ge, Zhiwei Jiang, Zifeng Cheng, Cong Wang, Yafeng Yin, and Qing Gu. 2023. Learning Robust Multi-Modal Representation for Multi-Label Emotion Recognition via Adversarial Masking and Perturbation. 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/10.1145/3543507.3583258

1 INTRODUCTION

Multi-modal Multi-label Emotion Recognition (MMER) task aims to identify emotions from multi-modal data, so as to meet the requirements of some special applications, such as chatbot and emotional conversation systems [30]. Compared with emotion recognition based on single modality, such as text-based emotion recognition, MMER requires stronger multi-modal representation ability. The key issue of improving multi-modal representation ability is how to effectively extract representation from each single modality and integrate representations of all modalities.

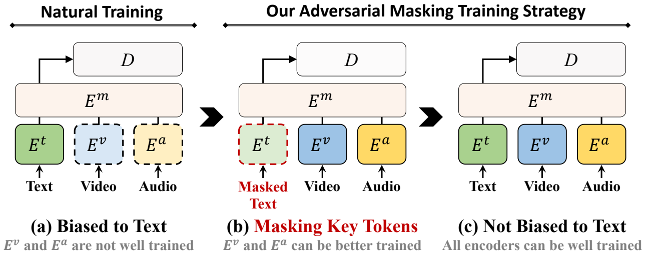
Existing MMER methods generally focus on how to design effective network structures to obtain strong representation ability, and just naturally train the model on training data without intervention [1, 10, 13, 31]. They expect that natural training can allow the representation model freely and adaptively learn the most relevant information from the training data for emotion recognition.

However, despite the good performance they achieve, the natural training scheme is prone to encounter two problems. Firstly, natural training does not guarantee that every modality can be adequately encoded. As shown in the left part of Figure 1, it is possible that the model tends to over-encode some informative modality (such as the text modality in Figure 1) while under-encode other less

[*] Corresponding author.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583258



(a) Biased to Text
E^v and E^a are not well trained

(b) Masking Key Tokens
E^v and E^a can be better trained

(c) Not Biased to Text
All encoders can be well trained

Figure 1: The illustration of natural training and our proposed adversarial masking training strategy.

informative modalities. We denote this problem as modality bias of representation. Such modality bias may result in the unstable performance of model on unseen data when the under-encoded modality is dominant for recognition [26]. Secondly, natural training without intervention (e.g., regularization) may cause the model overfit the training data. We denote this problem as data bias of training. Such training data bias may lead to the weak generalization of model on unseen data with different data distribution.

To address these problems, we propose a robust multi-modal representation learning method for multi-label emotion recognition, which advocates the use of adversarial training for multi-model representation learning. Specifically, we propose two adversarial training strategies. Firstly, to address the modality bias, we propose an adversarial 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. As shown in the right part of Figure 1, when we mask the emotion-related information in text modality, the encoder of other modalities can be enhanced and the emotion-related information in other two modalities can be adequately encoded. The key challenge of this strategy is how to locate the most emotion-related units for masking, which can be solved by mask prediction and adversarial gradient reversal. Secondly, to address the training data bias, we propose an adversarial parameter perturbation strategy, aiming at enhancing the generalization of the model. Unlike traditional adversarial perturbation strategies that add perturbation to the input image pixels or word embeddings, we add the adversarial perturbation to the intermediate parameters of model as model regularization. The key challenge of this strategy is how to decide the perturbation for each parameter, which can be solved by perturbation in the opposite direction of parameter’s gradient.

The major contributions of this paper are summarized as follows:

- We design a simple encoder-decoder style multi-modal emotion recognition model, and combine it with our specially-designed adversarial training strategies to learn more robust multi-modal representation for multi-label emotion recognition.
- We propose two effective adversarial training strategies, i.e., adversarial temporal masking and adversarial parameter perturbation, which can better boost model performance than natural training.
- We conduct experiments on the benchmark datasets CMU-MOSEI and NEMU. The experimental results demonstrate that our proposed method outperforms previous methods and achieves state-of-the-art performance.

2 RELATED WORK

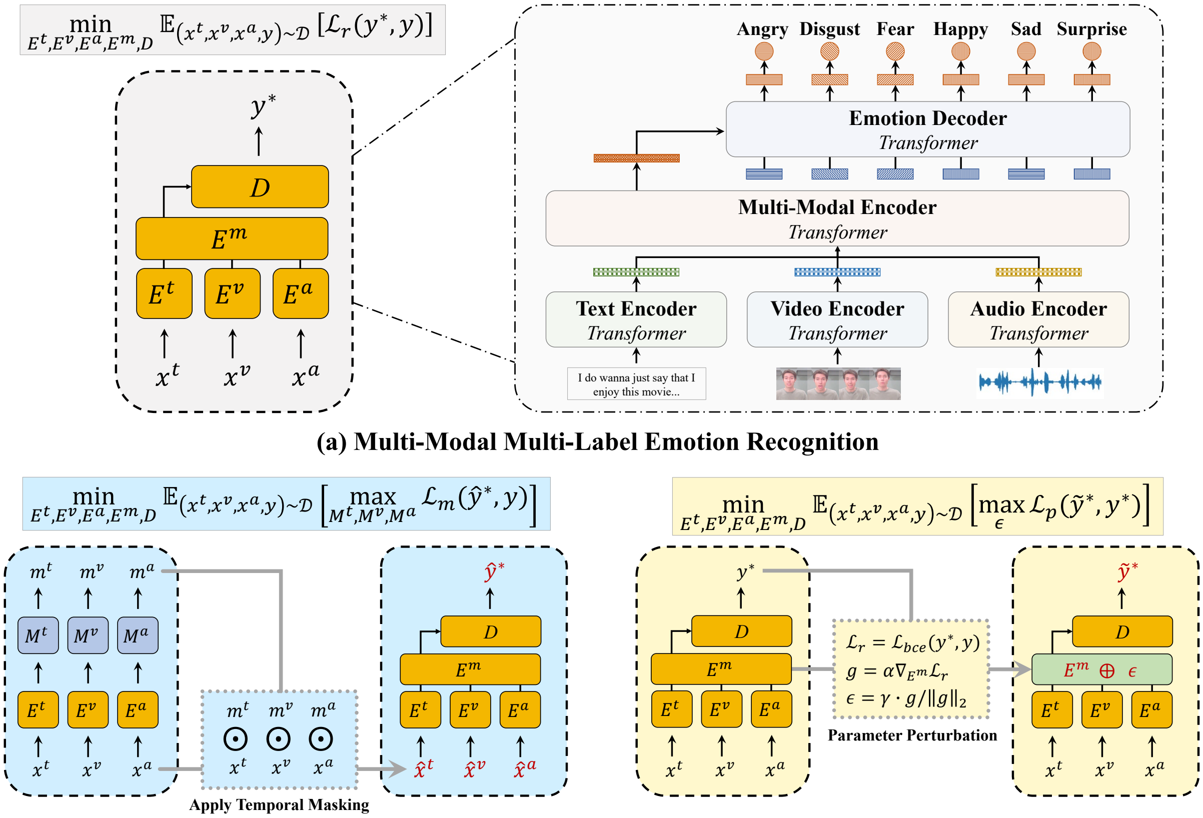
In this section, we introduce the following research topics relevant to our work comprehensively.

2.1 Multi-Modal Emotion Recognition

Multi-Modal Emotion Recognition is a research hotspot that has attracted widespread attention in the affective computing community. Previous methods mainly focus on the alignment and fusion of multi-modal data [5, 13, 25, 33, 34]. Tsai et al. [25] introduce the Multi-modal Transformer to generically address the issues of inherent data non-alignment and long-range dependencies across modalities. Chauhan et al. [5] learn the inter-modal interaction among different modalities with an auto-encoder. Ju et al. [13] use a transformer-based architecture to model the modality-to-label and label-to-label dependency simultaneously. More recently, a few methods for the more realistic MMER task have been proposed. Zhang et al. [31] introduce a multi-modal sequence-to-set approach to model the dependence in different labels and modalities. Zhang et al. [32] propose a heterogeneous hierarchical message passing network that considers label-to-label, feature-to-label, and modality-to-label dependency during training. Zhang et al. [33] devise a BERT-like cross-modal encoder to fuse private and common modality representations to capture richer semantic information for each label from different perspectives. While these methods only focus on the design of model structures, they neglect the problems of data bias and modality bias during training.

2.2 Bias in Multi-Modal Learning

Recent studies have demonstrated that the modality bias problem is often encountered in multi-modal learning, and many methods have been proposed to tackle this problem [7, 16, 24, 26]. Winterbottom et al. [26] demonstrate an inherent bias in the dataset towards the textual subtitle modality in the large-scale video question answering dataset TVQA. Gat et al. [7] propose a novel regularization term based on the functional entropy to balance the contribution of each modality to the multi-modal classification result. Tian et al. [24] discover that the discriminative information is not well-explored during training due to the modality bias problem and mitigate modality bias with an individual-guided learning mechanism. Liu et al. [16] find that existing models tend to capture the selection biases of frequently appeared video-query pairs in the temporal sentence grounding task and propose the feature distillation and contrastive sample generation methods to alleviate the issue. Besides modality bias, the data bias problem, a common problem in supervised learning, can also be encountered in multi-modal learning. Previous studies have tried to alleviate the data bias problem through various strategies, including feature selection [9], data preprocessing [2], and model adjustment [14]. While these methods are designed for uni-modal learning, they are also applicable for multi-modal learning.


$\min_{E^t,E^v,E^a,E^m,D} \mathbb{E}_{(x^t,x^v,x^a,y)\sim\mathcal{D}}[\mathcal{L}_r(y^*,y)]$
y^*
D
E^m
E^t
E^v
E^a
x^t
x^v
x^a
Angry
Disgust
Fear
Happy
Sad
Surprise
Emotion Decoder
Transformer
Multi-Modal Encoder
Transformer
Text Encoder
Transformer
Video Encoder
Transformer
Audio Encoder
Transformer
I do wanna just say that I enjoy this movie...
$\min_{E^t,E^v,E^a,E^m,D} \mathbb{E}_{(x^t,x^v,x^a,y)\sim\mathcal{D}}\left[\max_{M^t,M^v,M^a} \mathcal{L}_m(\hat{y}^*,y)\right]$
m^t m^v m^a
M^t M^v M^a
$m^t \odot x^t$ $m^v \odot x^v$ $m^a \odot x^a$
Apply Temporal Masking
$\hat{y}^*$
$\hat{x}^t$ $\hat{x}^v$ $\hat{x}^a$
$\min_{E^t,E^v,E^a,E^m,D} \mathbb{E}_{(x^t,x^v,x^a,y)\sim\mathcal{D}}\left[\max_{\epsilon} \mathcal{L}_p(\tilde{y}^*,y^*)\right]$
$\mathcal{L}_r = \mathcal{L}_{bce}(y^*,y)$
$g = \alpha\nabla_{E^m}\mathcal{L}_r$
$\epsilon = \gamma \cdot g/\|g\|_2$
Parameter Perturbation
$\tilde{y}^*$
$E^m \oplus \epsilon$

(a) Multi-Modal Multi-Label Emotion Recognition

(b) Adversarial Temporal Masking

(c) Adversarial Parameter Perturbation

Figure 2: The illustration of our MMER model and proposed two adversarial training strategies.

2.3 Adversarial Training

Adversarial Training (AT) focuses on revealing the defects of models and improving the robustness [8, 15]. The main process of adversarial training is to inject adversarial examples designed by an adversary to improve the robustness of the model [3, 4, 29]. Adversarial training has been explored in many CV and NLP tasks. Kurakin et al. [15] use adversarial training to train a model on the large-scale ImageNet dataset, significantly improving the model's robustness. Miyato et al. [19] extend adversarial training to the text domain by applying perturbations to the word embeddings in a RNN model to learn a robust text classification model. Wu et al. [27] apply AT in relation extraction within the multi-instance multi-label learning framework. Shafahi et al. [22] present an algorithm that eliminates the overhead cost of generating adversarial examples by recycling the gradient information computed when updating model parameters. In this paper, we use the idea of adversarial training to alleviate the data and modality bias and learn more robust multi-modal representation for the MMER task.

3 PROPOSED METHOD

In this section, we first define the multi-modal multi-label emotion recognition task formally and then introduce the model architecture and the proposed two adversarial training strategies. Finally, we give the overall training algorithm in detail.

3.1 Task Definition

We first introduce notations and formalize the Multi-modal Multi-label Emotion Recognition (MMER) task. Without loss of generality, we consider three modalities, i.e., text modality, video modality, and audio modality, in the MMER task. Let $\mathcal{Y} = \{1, 2, \ldots, K\}$ denote the pre-defined emotion label set with K different emotion labels. Let $\mathcal{D} = \{(\boldsymbol{x}_i^t, \boldsymbol{x}_i^v, \boldsymbol{x}_i^a), \boldsymbol{y}_i\}_{i=1}^N$ denote the training set, where $(\boldsymbol{x}_i^t, \boldsymbol{x}_i^v, \boldsymbol{x}_i^a)$ is the i-th multi-modal sample, $\boldsymbol{y}_i \subseteq \mathcal{Y}$ is the ground-truth label set of the i-the sample, and N is the number of samples in the training set. Note that each sample contains three sequences of temporal units (i.e., words for text, frames for video, and segments for audio) corresponding to three modalities, i.e., text sequence $\boldsymbol{x}_i^t = \{\boldsymbol{x}_{ij}^t\}_{j=1}^{L^t}$, video sequence $\boldsymbol{x}_i^v = \{\boldsymbol{x}_{ij}^v\}_{j=1}^{L^v}$, and audio sequence $\boldsymbol{x}_i^a = \{\boldsymbol{x}_{ij}^a\}_{j=1}^{L^a}$, where L^t, L^v, and L^a denote the length of corresponding sequence. Then, the goal of the MMER task is to learn a model based on training set $\mathcal{D}$ and apply it to recognize emotions from unseen samples.

3.2 Overview of Model Architecture

We design an encoder-decoder style model for the MMER task. As shown in Figure 2(a), our model consists of five Transformer-based modules: text encoder E^t, video encoder E^v, audio encoder E^a, multi-modal encoder E^m, and emotion decoder D. For convenience,

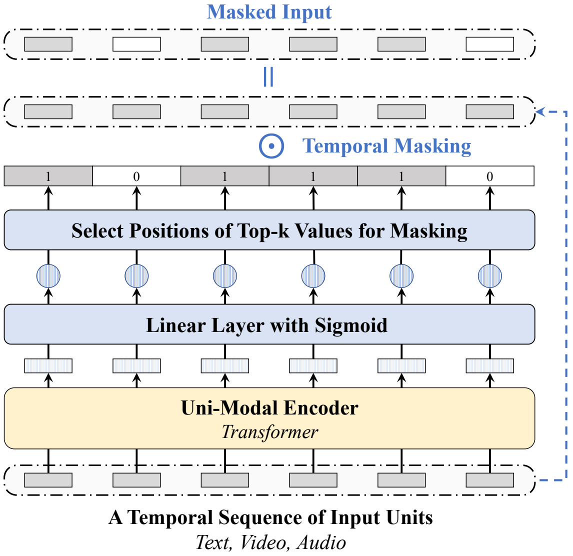

Masked Input
||
Temporal Masking
1
0
1
1
1
0
Select Positions of Top-k Values for Masking
Linear Layer with Sigmoid
Uni-Modal Encoder
Transformer
A Temporal Sequence of Input Units
Text, Video, Audio

Figure 3: The detailed illustration of ATM strategy.

we denote all encoders as $E = \{E^t, E^v, E^a, E^m\}$ and the parameters of our model as $\theta = \{\theta_E, \theta_D\}$, where θ_E refers the parameters of E.

Given a multi-modal sample $(\boldsymbol{x}^t, \boldsymbol{x}^v, \boldsymbol{x}^a)$ as input, we first use three uni-modal encoders E^t, E^v, and E^a to extract features $\boldsymbol{H}^t$, $\boldsymbol{H}^v$, and $\boldsymbol{H}^a$ from $\boldsymbol{x}^t$, $\boldsymbol{x}^v$, and $\boldsymbol{x}^a$, respectively. Then, we concatenate these features into a unified feature vector $\boldsymbol{H}^c = [\boldsymbol{H}^t, \boldsymbol{H}^v, \boldsymbol{H}^a]$, and sent it to E^m to get the multi-modal representation $\boldsymbol{H}^m$. After that, we set K learnable label embeddings as queries of emotion decoder to decode $\boldsymbol{H}^m$. Finally, the outputs of D are fed to K linear classifiers with Sigmoid function to obtain the probabilities $\boldsymbol{y}^*$ of all emotions in $\mathcal{Y}$.

Therefore, the optimization objective of training the model on the training set $\mathcal{D}$ can be written as:

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

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

where $\mathcal{L}_{bce}(\cdot)$ refers the binary cross-entropy loss function:

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

3.3 Adversarial Temporal Masking

The key idea of Adversarial Temporal Masking (ATM) strategy is to mask the most emotion-related temporal units of one modality, thus boosting the representation ability of other modalities. Such masking operation can be applied to all modalities simultaneously. Therefore, the critical challenge is how to identify the most emotion-related temporal units of each modality. As shown in Figure 2(b), we solve this problem by setting a masking network M for each modality (i.e., M^t, M^v, and M^a for text, video, and audio, respectively) to predict the probability of each temporal unit of input sequence to be emotion-related. Taking M^t as an example, the masking network M^t can take the output H^t of encoder E^t as input, and then generate a binary mask $\boldsymbol{m}^t$ with the same length of the input sequence $\boldsymbol{x}^t$. Then, by applying $\boldsymbol{m}^t$ to the original input, we can obtain a masked input $\hat{\boldsymbol{x}}$. Finally, we can get the recognition probabilities $\hat{\boldsymbol{y}}^*$ of the masked input $\hat{\boldsymbol{x}}$.

To ensure that the masking network can be well optimized and correctly predict the emotion-related temporal units, the masking network should play a minmax game with the encoder E and decoder D, which can be formalized as:

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

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

which implies that all masking networks M^t, M^v, and M^a try to mask all emotion-related units, thus making the recognition model unable to correctly predict the emotions (i.e., maximizing the loss), while encoder and decoder expect to correctly predict the emotions (i.e., minimizing the loss).

Specifically, as shown in Figure 3, we use a linear layer with Sigmoid function and a top-k operator to construct the masking network M. Given an input sequence $\boldsymbol{x}$, the binary mask $\boldsymbol{m}$ can be calculated as:

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

where $\boldsymbol{H} \in \mathbb{R}^{L \times d}$ is the output of uni-modal encoder for $\boldsymbol{x}$, $\boldsymbol{W} \in \mathbb{R}^{d \times 1}$, $\boldsymbol{b} \in \mathbb{R}^{L \times 1}$ are the weight and bias of the linear layer respectively, L is the length of input sequence $\boldsymbol{x}$, and d is the dimension of feature vector H. Let $\boldsymbol{s} \in \mathbb{R}^{L \times 1}$ denote a real-value sequence, the top-k operator $\mathcal{T}$ is defined as:

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

In order to enable the parameters of M to be optimized by back-propagation, we employ the SOFT top-k operator proposed in [28] as the top-k operator instead. The SOFT top-k operator parameterizes the standard top-k operator in terms of the solution of the following Optimal Transport problem:

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

where Γ^* is the optimal transport plan, $C \in \mathbb{R}^{n \times m}$ is the cost matrix using squared Euclidean distance. Then the top-k operator can be rewritten as:

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

After that, the masked input $\hat{\boldsymbol{x}}$ can be calculated by:

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

where $\odot$ represents the temporal masking operation which masks the temporal units of $\boldsymbol{x}$ according to the value of $\boldsymbol{m}$ at the same position.

In addition, to optimize the masking network based on the minmax equation (i.e., Eq.4), we should reverse the gradient of M during back-propagation, so that the model can be penalized when masking temporal units with less emotion-relation information during training. The optimization formula of the parameters of M is:

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

where α is the learning rate.

Algorithm 1 Overall Training Flow of the Proposed Adversarial Training Strategies

```
Input: The training set $\mathcal{D} = \{(\boldsymbol{x}_i^t, \boldsymbol{x}_i^v, \boldsymbol{x}_i^a), \boldsymbol{y}_i\}_{i=1}^N$ .
Output: The well-trained model.
Initialize model parameters $\theta_E$, $\theta_D$, $\theta_{M^t}$, $\theta_{M^v}$, $\theta_{M^a}$, and learning rate $\alpha$.
Repeat
    ▷ Emotion Recognition:
    $\boldsymbol{y}^* = D(E^m(E^t(\boldsymbol{x}^t), E^v(\boldsymbol{x}^v), E^a(\boldsymbol{x}^a)))$.
    $\mathcal{L}_r = \mathcal{L}_{bce}(\boldsymbol{y}^*, \boldsymbol{y})$.
    Compute $\nabla_{\theta_{E^t}} \mathcal{L}_r$, $\nabla_{\theta_{E^v}} \mathcal{L}_r$, $\nabla_{\theta_{E^a}} \mathcal{L}_r$, $\nabla_{\theta_{E^m}} \mathcal{L}_r$, and $\nabla_{\theta_D} \mathcal{L}_r$ through back-propagation.
    ▷ Adversarial Temporal Masking:
    Compute masks: $\boldsymbol{m}^t = M^t(E^t(\boldsymbol{x}^t))$, $\boldsymbol{m}^v = M^v(E^v(\boldsymbol{x}^v))$, $\boldsymbol{m}^a = M^a(E^a(\boldsymbol{x}^a))$.
    $\hat{\boldsymbol{x}}^t = \boldsymbol{x}^t \odot \boldsymbol{m}^t$, $\hat{\boldsymbol{x}}^v = \boldsymbol{x}^v \odot \boldsymbol{m}^v$, $\hat{\boldsymbol{x}}^a = \boldsymbol{x}^a \odot \boldsymbol{m}^a$.
    $\hat{\boldsymbol{y}}^* = D(E^m(E^t(\hat{\boldsymbol{x}}^t), E^v(\hat{\boldsymbol{x}}^v), E^a(\hat{\boldsymbol{x}}^a)))$.
    $\mathcal{L}_m = \mathcal{L}_{bce}(\hat{\boldsymbol{y}}^*, \boldsymbol{y})$.
    Compute $\nabla_{\theta_{E^t}} \mathcal{L}_m$, $\nabla_{\theta_{E^v}} \mathcal{L}_m$, $\nabla_{\theta_{E^a}} \mathcal{L}_m$, $\nabla_{\theta_{E^m}} \mathcal{L}_m$, $-\nabla_{\theta_{M^t}} \mathcal{L}_m$, $-\nabla_{\theta_{M^v}} \mathcal{L}_m$, and $-\nabla_{\theta_{M^a}} \mathcal{L}_m$, through back-propagation.
    ▷ Adversarial Parameter Perturbation:
    Compute the perturbation $\epsilon_{E^m} = \gamma \cdot \frac{g}{\|g\|_2}$, and $g = \alpha \nabla_{\theta_{E^m}} \mathcal{L}_r$, and then apply it to $E^m$ to obtain $E^m_\epsilon$.
    $\tilde{\boldsymbol{y}}^* = D(E^m_\epsilon(E^t(\boldsymbol{x}^t), E^v(\boldsymbol{x}^v), E^a(\boldsymbol{x}^a)))$.
    $\mathcal{L}_p(\tilde{\boldsymbol{y}}^*, \boldsymbol{y}^*) = \mathrm{KL}(p(\boldsymbol{y}^* | \boldsymbol{x}^t, \boldsymbol{x}^v, \boldsymbol{x}^a; \theta) \| p(\tilde{\boldsymbol{y}}^* | \boldsymbol{x}^t, \boldsymbol{x}^v, \boldsymbol{x}^a; \theta + \epsilon))$.
    Compute $\nabla_{\theta_{E^t}} \mathcal{L}_p$, $\nabla_{\theta_{E^v}} \mathcal{L}_p$, $\nabla_{\theta_{E^a}} \mathcal{L}_p$, $\nabla_{\theta_{E^m}} \mathcal{L}_p$, and $\nabla_{\theta_D} \mathcal{L}_p$ through back-propagation.
    ▷ Optimization:
    Accumulate all the gradients based on the tradeoff parameter in Eq.15 and optimize the model with AdamW optimizer.
Until model converges or reaches maximum iterations.
```

3.4 Adversarial Parameter Perturbation

The key idea of Adversarial Parameter Perturbation (APP) strategy is to obtain noise-invariant representation by perturbing the model parameters without affecting the prediction results, thus enhancing the generalization of the model. As is shown in Figure 2(c), unlike traditional adversarial training methods that add perturbation to the input image pixels or word embeddings for various uni-modal tasks, we apply adversarial perturbation ϵ to the intermediate parameters of the model to work as an effective regularization during the learning of multi-modal representations.

To find the effective perturbation ϵ, the perturbation also should play a minmax game with the encoder E and decoder D, which can be formalized as:

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

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

where γ is the constraint for the perturbation, $\tilde{\boldsymbol{y}}^*$ is the output of the perturbed model, and $\mathrm{KL}(p \| q)$ denotes the KL divergence between distributions p and q. This minmax equation implies that all perturbation ϵ try to perturb the model parameters, thus making the recognition model unable to correctly predict the emotions (i.e., maximizing the loss), while encoder and decoder expect to resist the perturbation and correctly predict the emotions (i.e., minimizing the loss).

Specifically, we perturb the parameters of the multi-modal encoder E^m by adding learnable noise $\mathrm{Attention}_\epsilon$ and FNN_ϵ to the projection layers of attention mechanism and the feed-forward layers during training, respectively.

However, exact minimization of ϵ is intractable for neural networks. Inspired by [8], we approximate ϵ by linearizing $\mathcal{L}_r$ around input $\boldsymbol{x}$. With a linear approximation and an L_2 norm constraint, the perturbation is calculated every training step as:

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

During every training step, we first compute $\mathcal{L}_r$ and then back-propagate to get the gradient g and apply the noise ϵ to the multi-head self-attention layers of the multi-modal transformer encoder E^m, and then re-optimize the model again.

3.5 Overall Training

In summary, the overall training objective of our method is:

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

where ρ and σ are two trade-off parameters.

The full training flow of our model is provided in Algorithm 1. Given a training batch $\{(\boldsymbol{x}_i^t, \boldsymbol{x}_i^v, \boldsymbol{x}_i^a, \boldsymbol{y}_i)\}_{i=1}^B$ with B samples, first, we input the clean data into the model and use the output to compute the loss $\mathcal{L}_r$, then perform back-propagation to compute the gradient $\nabla_\theta \mathcal{L}_r$. Second, we use the masking networks M^t, M^v, and M^a to generate the mask to obtain the masked data $(\hat{\boldsymbol{x}}^t, \hat{\boldsymbol{x}}^v, \hat{\boldsymbol{x}}^a, \boldsymbol{y})$ and then compute $\mathcal{L}_m$ and $\nabla_\theta \mathcal{L}_m$ and reversed gradient $-\nabla_{\theta_M} \mathcal{L}_m$. Third, we use $\mathcal{L}_r$ to compute the perturbation ϵ and then compute $\mathcal{L}_p$ and $\nabla_\theta \mathcal{L}_p$. Finally, we update the model with the weighted accumulated gradients using the AdamW optimizer.

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.

4 EXPERIMENTS

In this section, we first introduce the datasets, evaluation metrics, and implementation details of our method. Then, we compare our method with other baseline methods on both datasets. Finally, we conduct ablation studies and model analysis for our method.

4.1 Experimental Setup

4.1.1 ***Dataset.*** To validate the effectiveness of our method, we conduct experiments on two large-scale multi-modal multi-label emotion datasets, i.e., CMU-MOSEI [30] and NEMu [32].

CMU-MOSEI The dataset contains 3,229 full-long videos from 1,000 distinct speakers, and the videos are segmented into 22,856 utterance-level video clips. Each utterance-level video clip contains data of three modalities, i.e., text, video, and audio, and is annotated with six discrete human emotions: *angry*, *disgust*, *fear*, *happy*, *sad*, and *surprise*. The features are pre-extracted following the previous work [30, 33]. For CMU-MOSEI, the 300-D text features are from the video transcripts and are extracted using GloVe model [21], the 35-D video features are extracted from commonly used facial recognition models DeepFace [23], and the 74-D audio features are extracted using the COVAREP software [6].

NEMu NEMu is a partial time series dataset collected from the streaming music platform NetEase Cloud Music. Each sample in this dataset contains lyrics, comments, audios, and images (e.g., covers and posters), and is annotated with twelve discrete human emotions including *sad*, *excited*, *lonely*, *quiet*, etc. The features are pre-extracted following the previous work [32]. For NEMu, the 300-D lyric and comment features are extracted using GloVe model, the 74-D audio features are extracted using Librosa for MFCCs, and the 2048-D image features are extracted using ResNet model [11].

4.1.2 ***Evaluation Metrics.*** To make a fair comparison with previous methods, we employ four widely-used evaluation metrics to measure the performance of different methods on the multi-label classification problems, i.e., multi-label Accuracy (Acc), Hamming Loss (HL), micro F_1 (miF_1), and macro F_1 (maF_1). Larger Acc, miF_1, maF_1 and smaller HL indicate better recognition performance.

4.1.3 ***Implementation Details.*** We implement our model using Pytorch [20] and train on 2 GPUs. For the encoder-decoder model, we set the number of transformer blocks as 4 for both uni-modal and multi-modal encoders, and 1 for the emotion decoder. The number of attention heads, dimension of hidden states, and feed-forward layers are set to 8, 512, and 1,024 in all transformer blocks, respectively. The complexity of our model is comparable to previous transformer-based methods. During training, we train the model for 10 epochs and choose the model that obtains the best Micro-F1 score on the validation set as the final model. We adopt the AdamW optimizer [17] with an initial learning rate of 5e-5 and batch size of 64 to optimize the model. Additionally, the cosine annealing scheduler is adopted to adjust the learning rate and the warm-up strategy is performed at the first 10 iterations. We set the hyper-parameters $\alpha = 0.1$, $\rho = 0.1$, and $\sigma = 1$.

4.2 Comparing with Existing Methods

Considering that there are only a few multi-modal multi-label emotion recognition methods for comparison, we also compare our method with various uni-modal multi-label (classical, linguistic, and non-linguistic) and multi-modal single-label emotion recognition methods, following the setting in previous studies [32]. For uni-modal multi-label methods, we replace their input uni-modal features with multi-modal features (i.e., the features from different modalities are early-fused to be a unified multi-modal feature). For multi-modal single-label methods, we replace their final multi-class classification layer with the multi-label classification layer (i.e., a linear layer with the Sigmoid activation function).

Table 1 presents the comparison results of our method and the related methods on both CMU-MOSEI and NEmu datasets. From

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

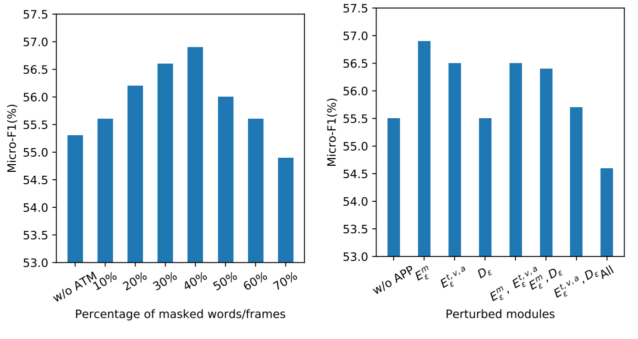
Table 2: Ablation study of our model. Accuracy, Precision, Recall, and Micro-F1 scores on the CMU-MOSEI dataset and NEMu dataset. '$-E^t$ and x^t' means removing the text encoder and text input. Note that for NEMu, $-E^t$ means removing lyrics and comments features at the same time.

the results we can see that: First, the linguistic and non-Linguistic multi-label classification methods achieve better performance than classical machine learning methods on the CMU-MOSEI dataset, while the results are opposite on the NEmu dataset generally. For example, LASN performs worse than CC on the NEMu dataset, but obviously outperforms CC on the CMU-MOSEI dataset. The results suggest that multi-modal data requires high content analysis ability of the model since the difference of the source and organization of the multi-modal datasets could cause significant influence in experimental results. Second, the multi-modal methods achieve better results than uni-modal methods, which demonstrates the effectiveness and necessity of leveraging multi-modal information for recognizing human emotions and multi-modal data need to well model interactions among different modalities. Third, our method outperforms previous methods on the Acc, HL, miF_1 metrics on CMU-MOSEI, and Acc, miF_1, maF_1 metrics on NEMu, which validates the effectiveness and generalization ability of our method.

4.3 Ablation Study

In this part, we explore the effects of the components designed in our method by removing each of them individually.

4.3.1 ***Effect of Model Architecture.*** We conduct ablation study on the model components to validate the effectiveness of model architecture. These results are obtained by first ablating components from model and then training model with ATM and APP. Specifically, we remove the text encoder E^t, video encoder E^v, audio encoder E^a, multi-modal encoder E^m (replaced by concatenating output of different uni-modal encoders), emotion decoder D (replaced by a linear classifier after E^m), and transformer architecture of the encoders (replaced by LSTM) separately. As shown in Table 2, removing different uni-modal encoders all decrease the recognition performance and removing text encoder E^t along with text data caused the greatest performance degradation. Moreover, the recognition performance degrades when the multi-modal encoder E^m and emotion decoder D are removed respectively, which shows the ability of multi-modal fusion and extraction of our model.

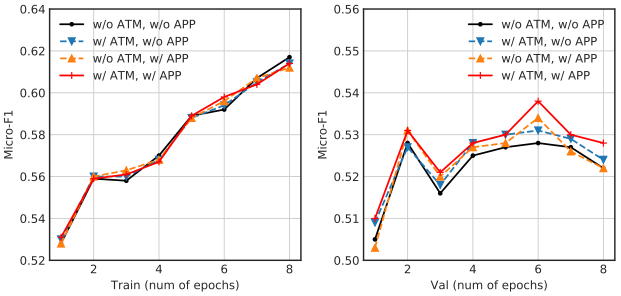


(a)

(b)

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. $E^{t,v,a}_\epsilon$ denotes applying APP to the uni-modal encoders E^t, E^v, and E^a.

4.3.2 ***Effect of Adversarial Temporal Masking.*** To study the effectiveness of ATM, we compare the ATM strategy with following variants: 1. ATM strategy is removed. 2. Remove the masking



(a) (b)

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

networks M^t, M^v, and M^a, then mask units randomly. 3. Optimize the masking networks without reversing the gradient $\nabla_{\theta_M}\mathcal{L}_m$. As shown in Table 2, removing the ATM strategy or part of ATM, the performance of model variants shows varying degrees of degradation. The results indicate the modality bias in the natural training, and imply the effectiveness of ATM and the importance of the masking networks and reversing the gradient $\nabla_{\theta_M}\mathcal{L}_m$ for mining the rich semantic information during training.

4.3.3 *Effect of Adversarial Parameter Perturbation.*

To study the effectiveness of APP, we compare our APP strategy with following variants: 1. APP strategy is removed. 2. Replace gradient-based perturbation with random perturbation. 3. Remove the perturbation applied on the attention mechanism. 4. Remove the perturbation applied on the feed-forward layers. As shown in Table 2, the recognition performance degrades on all four evaluation metrics without the APP strategy, especially on P. Moreover, replacing the approximation method of perturbation with random perturbation based on normal distribution performs even worse compared to the model without APP strategy, which show the importance of the approximation solution and constraint of the perturbation.

Furthermore, in the last five rows of Table 2, we show the results of removing ATM and APP at the same time, and replace the adversarial training strategies with previous adversarial training methods FGSM [8] and PGD [18] and basic data augmentation methods random erasing [35] and Gaussian noise [12]. It can be seen from the results that traditional adversarial training and data augmentation strategies have no obvious effect on the MMER task.

4.4 Model Analysis

In this part, we analyze the impacts of the masked units in ATM and the perturbed modules in APP on the performance of our model, and the robustness of our model.

4.4.1 *Impact of the Numbers of Masked Units in ATM.*

We conduct experiments to investigate the influence of different numbers of masked units in ATM during training. We set the percentages of masked words/tokens from 10% to 80% step by 10% and then train the model with different percentages. The results are shown in Figure 4 (a). It can be seen from the results that masking 40% during training achieves the best results, while masking excessive units could lead to performance degradation.

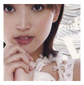
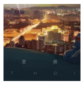
Images	***Lyrics***	***Comments***	**Ground-Truth**	**Ours**	**HHMPN**
	Imagine if you can meet again at this moment, will you forget the past...	*I never thought that so many friends have heard this song, and it is very touching...*	*Sad* *Lonely* *Miss* *Quiet* *Healing*	*Sad* *Lonely* *Miss* *Healing*	*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.

4.4.2 *Impact of Perturbing Different Modules in APP.*

Furthermore, we conduct experiments on applying APP to the attention mechanism and feed-forward layers in the different modules of the model. We consider adding perturbation to the uni-modal encoders E^t, E^v, E^a, the emotion decoder D, and different combinations of these modules. It can be observed from Figure 4 (b) that applying APP to the multi-modal encoder E^m outperforms all other settings, and applying APP to all the modules could degrade the performance. The results show the effectiveness of perturbing the multi-modal encoder and excessive perturbation may deteriorate the optimization of the model.

4.4.3 *Robustness of Our Model.*

We finally validate whether our adversarial training strategies can make the model more robust and generalized on unseen data. First, we compare the performance of model with natural training (i.e., without ATM and APP) and with our adversarial training (i.e., with ATM, with APP, or with both ATM and APP) on validation set. As shown in Figure 5, as the performance on training set increases, the performance on the validation set starts to decrease rather than continue to increase. This phenomenon is called overfitting. Whereas, on the validation set, we can find that the line of using ATM and APP is always above the line of natural training. This indicates that our adversarial training strategies can alleviate the overfitting and provide better and more stable performance than natural training on unseen data (e.g., validation set). Second, we also show some recognition cases in the test set of NEMu in Figure 6. It can be seen from the cases that our method can make more positive predictions and fewer negative predictions than HHMPN. This also indicates the good generalization ability of our method on unseen data.

5 CONCLUSION

In this paper, we propose a novel method to learn robust multimodal representation for MMER based on adversarial training. We first propose a simple yet effective Transformer-based Encoder-Decoder model for emotion recognition, and then train it with our proposed two adversarial training strategies ATM and APP. Experimental results on the benchmark MMER datasets CMU-MOSEI and NEMu demonstrate that both strategies can boost the performance and generalization of the model, and the proposed method can outperform previous state-of-the-art method.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grants Nos. 61972192, 62172208, 61906085, 41972111. This work is partially supported by Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[2] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[3] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

[4] Shravan Chandra and Bhaskarjyoti Das. 2022. An approach framework of transfer learning, adversarial training and hierarchical multi-task learning-a case study of disinformation detection with offensive text. In *Journal of Physics: Conference Series*, Vol. 2161. IOP Publishing, 012049.

[5] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5647–5657.

[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] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems* 33 (2020), 3197–3208.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[9] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[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] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Liang Huang, Weijian Pan, You Zhang, Liping Qian, Nan Gao, and Yuan Wu. 2019. Data augmentation for deep learning-based radio modulation classification. *IEEE access* 8 (2019), 1498–1506.

[13] Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of the 28th ACM International Conference on Multimedia*. 512–520.

[14] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 35–50.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).

[16] Daizong Liu, Xiaoye Qu, and Wei Hu. 2022. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4092–4101.

[17] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[19] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725* (2016).

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[21] 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.

[22] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems* 32 (2019).

[23] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1701–1708.

[24] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 436–454.

[25] 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.

[26] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. 2020. On modality bias in the TVQA dataset. *arXiv preprint arXiv:2012.10210* (2020).

[27] Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1778–1783.

[28] Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. 2020. Differentiable top-k with optimal transport. *Advances in Neural Information Processing Systems* 33 (2020), 20520–20531.

[29] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems* 30, 9 (2019), 2805–2824.

[30] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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.

[31] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3584–3593.

[32] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14338–14346.

[33] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. *arXiv preprint arXiv:2201.05834* (2022).

[34] Sicheng Zhao, Yunsheng Ma, Yang Gu, Jufeng Yang, Tengfei Xing, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. 2020. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 303–311.

[35] 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.