CARAT: Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition

Cheng Peng, Ke Chen *, Lidan Shou, Gang Chen *

The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou 310000, China
{chengchng, chenk, should, cg}@zju.edu.cn

Abstract

Multi-modal multi-label emotion recognition (MMER) aims to identify relevant emotions from multiple modalities. The challenge of MMER is how to effectively capture discriminative features for multiple labels from heterogeneous data. Recent studies are mainly devoted to exploring various fusion strategies to integrate multi-modal information into a unified representation for all labels. However, such a learning scheme not only overlooks the specificity of each modality but also fails to capture individual discriminative features for different labels. Moreover, dependencies of labels and modalities cannot be effectively modeled. To address these issues, this paper presents ContrAstive feature Reconstruction and AggregaTion (CARAT) for the MMER task. Specifically, we devise a reconstruction-based fusion mechanism to better model fine-grained modality-to-label dependencies by contrastively learning modal-separated and label-specific features. To further exploit the modality complementarity, we introduce a shuffle-based aggregation strategy to enrich co-occurrence collaboration among labels. Experiments on two benchmark datasets CMU-MOSEI and M3ED demonstrate the effectiveness of CARAT over state-of-the-art methods. Code is available at https://github.com/chengzju/CARAT.

Introduction

Multi-modal Multi-label Emotion Recognition (MMER) aims to identify multiple emotions (e.g., happiness and sadness) from multiple heterogeneous modalities (e.g., text, visual, and audio). Over the last decades, MMER has fueled research in many communities, such as online chatting (Galik and Rank 2012), news analysis (Zhu, Li, and Zhou 2019) and dialogue systems (Ghosal et al. 2019).

Different from single-modal tasks, multi-modal learning synergistically processes heterogeneous information from various sources, which introduces a challenge of how to capture discriminative representations from multiple modalities. To this end, recent works propose various advanced multi-modal fusion strategies to bridge the modality gap and learn effective representations (Ramachandram and Taylor 2017). According to the fusion manner, methods can be roughly divided into three categories: aggregation-based, alignment-based, and the mixture of them(Baltrušaitis, Ahuja, and Morency 2019). The aggregation-based fusion employs averaging (Hazirbas et al. 2017), concatenation (Ngiam et al. 2011) or attention (Zadeh et al. 2018a) to integrate multi-modal features. The alignment-based fusion (Pham et al. 2018, 2019) adopts the cross-modal adaptation to align latent information of different modalities. However, unifying multiple modalities into one identical representation can inevitably neglect the specificity of each modality, thus losing the rich discriminative features. Although recent works (Hazarika, Zimmermann, and Poria 2020; Zhang et al. 2022) attempt to learn modality-specific representations, they still utilize attention to fuse these representations into one. Therefore, a key challenge of MMER is how to effectively represent multi-modal data while maintaining modality specificity and integrating complementary information.

Figure 1: An example of MMER (left) and correlations between two relevant emotions and the video sequence (right).

As a multi-label task (Zhang and Zhou 2013), MMER also needs to deal with complex dependencies among labels. Nowadays, massive studies attempt various methods to explore label correlation, such as label similarity (Xiao et al. 2019) and co-occurrence label graph (Ma et al. 2021). However, these static correlations cannot reflect the collaborative relationship among labels. On the other hand, another tricky conundrum for MMER is how to learn dependencies between labels and modalities. Commonly, different modalities have inconsistent emotional expressions, and conversely, different emotions focus on different modalities, which means that inferring each potential label largely depends on the different contributions of different modalities. As shown in Figure 1, we can infer *sadness* more easily from the visual modality, while *disgust* can be predicted from both textual and visual modalities. Therefore, another challenge

*Ke Chen and Gang Chen are the corresponding authors.

of MMER is how to effectively model both label-to-label and modality-to-label dependencies.

To address these issues, we propose ContrAstive feature Reconstruction and AggregaTion for MMER (CARAT), which coordinates representation learning and dependency modeling in a coherent and synergistic framework. Specifically, our framework CARAT encapsulates three key components. First, we adopt the label-wise attention mechanism to extract label-specific representations within each modality severally, which is intended to capture relevant discriminative features of each label while maintaining modality specificity. Second, to reconcile the complementarity and specificity of multi-modal information, we develop an ingenious reconstruction-based fusion strategy that attempts to generate features of any modality by exploiting the information from multiple modalities. We leverage contrastive learning (Khosla et al. 2020), which is unexplored in previous MMER literature, to facilitate the learning of modal-separated and label-specific features. Third, based on the reconstructed embeddings, we propose a novel sample-wise and modality-wise shuffle strategy to enrich the co-occurrence dependencies among labels. After shuffled, embeddings are aggregated to finetune a robust discriminator. Moreover, as for modeling the modality-to-label dependency, we employ a max pooling-like network to discover the most relevant modalities for different emotions per sample, and then impel these corresponding representations to be more discriminative. [1] The main contributions of this paper can be summarized as follows:

- A novel framework, ContrAstive feature Reconstruction and AggregaTion, is proposed. To the best of our knowledge, this work pioneers the exploitation of contrastive learning to facilitate a multi-modal fusion mechanism based on feature reconstruction. As an integral part of our method, we also introduce a shuffle-based feature aggregation strategy, which uses the reconstructed embeddings to better leverage multi-modal complementarity.
- To preserve the modality specificity, CARAT independently extracts label-specific representations from different modalities via label-wise attention. Then a max pooling-like network is involved to select the most relevant modal representation per emotion to explore potential dependencies between modalities and labels.
- We conduct experiments on two benchmark datasets CMU-MOSEI and M^{3}ED. The experimental results demonstrate that our proposed method outperforms previous methods and achieves state-of-the-art performance.

Related Works

Multi-modal Learning aims to build models that can process and relate information from multiple modalities (Baltrušaitis, Ahuja, and Morency 2019). A fundamental challenge is how to effectively fuse multi-modal information. According to the fusion manner, methods can be roughly divided into three categories: aggregation-based, alignment-based, and hybrid methods. Aggregation-based methods use concatenation (Ngiam et al. 2011), tensor fusion (Zadeh et al. 2017; Liu et al. 2018) and attention (Zadeh et al. 2018a) to combine multiple modalities, but suffer from the modality gap. To bridge the gap, alignment-based fusion (Pham et al. 2018, 2019) exploits latent cross-modal adaptation by constructing a joint embedding space. However, alignment-based fusion neglects the specificity of each modality, resulting in the omission of discriminative information.

Multi-label Emotion Recognition is a foundational multi-label (ML) task and ML approaches can be quickly applied. BR (Boutell et al. 2004) decomposes the ML task into multiple binary classification ones while ignoring label correlations. To exploit the correlations, LP (Tsoumakas and Katakis 2006), CC (Read et al. 2011) and Seq2Seq (Yang et al. 2018) are proposed. To further explore label relationships, recent works leverage reinforced approach (Yang et al. 2019), multi-task pattern (Tsai and Lee 2020), and GCN model (Chen et al. 2019b). Another important task is to learn effective label representations. To compensate for the inability of a single representation to capture discriminative information of all labels, recent works (Chen et al. 2019a,b) utilize label-specific representations to capture the most relevant features for each label, which has been successfully applied to many studies (Huang et al. 2016; Xiao et al. 2019).

Contrastive learning (CL) is an effective self-supervised learning technique (Li et al. 2021; Oord, Li, and Vinyals 2018; Hjelm et al. 2019) . CL aims to learn a discriminative latent space where similar samples are pulled together and dissimilar samples are pushed apart. Motivated by the successful application of CL in unsupervised learning (Oord, Li, and Vinyals 2018; He et al. 2020), Supervised Contrastive Learning (SCL) (Khosla et al. 2020) is devised to promote a series of supervised tasks. Recently, CL has been applied to multi-modal tasks to strengthen the interaction between features of different modalities (Zheng et al. 2022; Franceschini et al. 2022; Zolfaghari et al. 2021). However, there has been no exploration of contrastive learning on multi-modal tasks in the multi-label scenario.

Methodology

In this section, we describe our CARAT framework, which comprises three sequential components (in Figure 2).

Problem Definition

We define notations for MMER. Let $\mathcal{X}^t \in \mathbb{R}^{n_t \times d_t}$, $\mathcal{X}^v \in \mathbb{R}^{n_v \times d_v}$ and $\mathcal{X}^a \in \mathbb{R}^{n_a \times d_a}$ be the heterogeneous feature spaces for textual (t), visual (v) and acoustic (a) modality respectively, where n_m and d_m denotes the sequence length and modality dimension respectively ($m \in \{t, v, a\}$ is used to represent any modality). And $\mathcal{Y}$ is the label space with $\mathcal{C}$ labels. Given a training dataset $\mathcal{D} = \{(\boldsymbol{X}_i^{\{t,v,a\}}, \boldsymbol{y}_i)\}_{i=1}^N$, MMER aims to learn a function $\mathcal{F} : \mathcal{X}^t \times \mathcal{X}^v \times \mathcal{X}^a \mapsto \mathcal{Y}$ to predict relevant emotions for each video. Concretely, $\boldsymbol{X}_i^m \in \mathcal{X}^m$ are asynchronous coordinated utterance sequences and $\boldsymbol{y}_i = \{0, 1\}^{\mathcal{C}}$ is the multi-hot label vector, where sign $\boldsymbol{y}_{i,j} = 1$ indicates that sample i belongs to class j, otherwise $\boldsymbol{y}_{i,j} = 0$.

[1] This paper's complete version with technical appendices is available at https://arxiv.org/abs/2312.10201

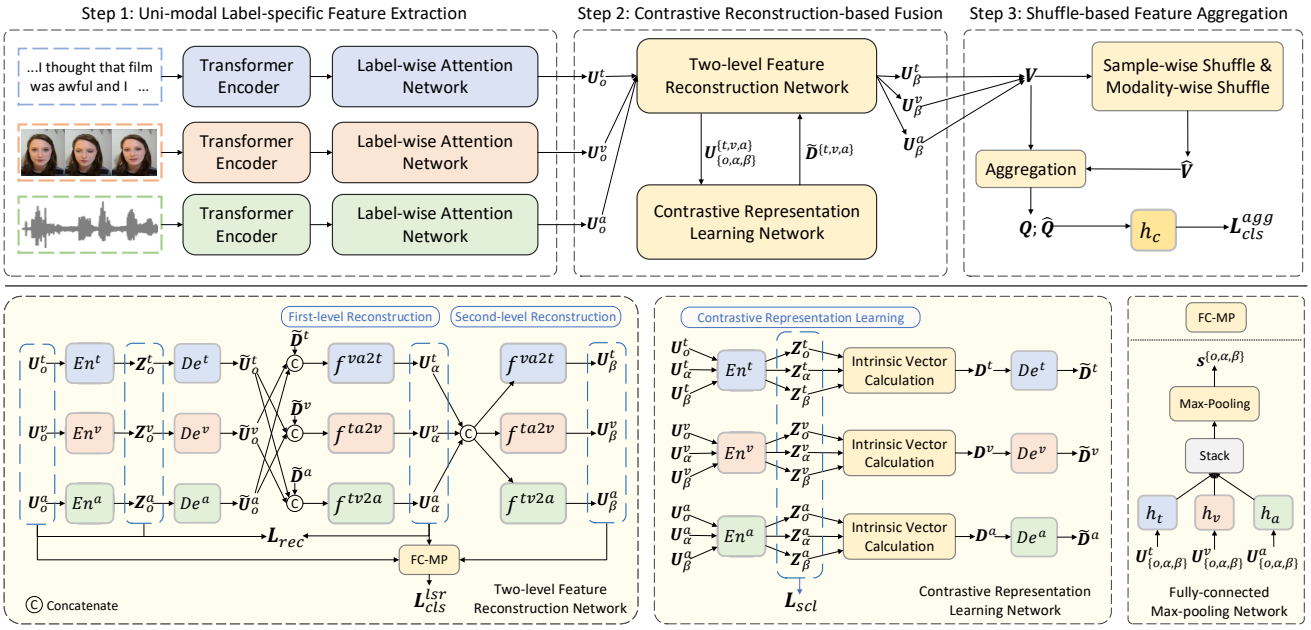

Figure 2: The overall structure of CARAT with three sequential steps (up). Detailed implementations of two-level feature reconstruction network, contrastive representation learning network, and Max Pooling-like network (bottom).

Uni-modal Label-specific Feature Extraction

As the first step, this component aims to extract the relevant discriminative features for each label in each modality.

Transformer-based Extractor. For each modality m, we use an independent Transformer Encoder (Vaswani et al. 2017) to map raw feature sequences $\boldsymbol{X}^m \in \mathbb{R}^{n_m \times d_m}$ into high-level embedding sequences $\boldsymbol{H}^m \in \mathbb{R}^{n_m \times d}$. Each encoder is composed of l_m identical layers, where each layer consists of two sub-layers: a multi-head self-attention sub-layer and a position-wise feed-forward sub-layer. The residual connection (He et al. 2016) is employed around each of the two sub-layers, followed by layer normalization.

Multi-label Attention. Considering that each emotion is usually expressed by the most relevant part of the utterance, we generate label-specific representations for each emotion to capture the most critical information. After obtaining embedding sequences $\boldsymbol{H}^m$, we compute the combination of these embeddings for each label j under each modality m through a label-wise attention network. Formally, we represent the hidden state of each embedding as $\boldsymbol{h}_i^m \in \mathbb{R}^d (i \in [n_m])$. The attentional representation $\boldsymbol{u}_j^m$ is obtained as:

$$\boldsymbol{u}_j^m = \sum_{i=1}^{n_m} \alpha_{ij}^m \boldsymbol{h}_i^m, \quad \alpha_{ij}^m = \frac{\exp(\boldsymbol{w}_j^{m\top} \boldsymbol{h}_i^m)}{\sum_{i'=1}^{n_m} \exp(\boldsymbol{w}_j^{m\top} \boldsymbol{h}_{i'}^m)}, \quad (1)$$

where $\boldsymbol{w}_j^m \in \mathbb{R}^d$ denotes the attention parameter for the j-th label and α_{ij}^m is the normalized coefficient of $\boldsymbol{h}_i^m$. It is worth noting that attention networks between modalities are still independent of each other, thus generating label-specific representations $\boldsymbol{U}_o^m \in \mathbb{R}^{\mathcal{C} \times d}$ separately.

Contrastive Reconstruction-based Fusion

The second component aims to utilize information from multiple modalities to restore the features of any modality.

Multi-modal Feature Reconstruction. Considering that fusing multi-modal information into an identical representation can ignore the modality specificity, we propose a reconstruction-based fusion mechanism, which restores features of any modality with the feature distribution in the current modality and the semantic information in other modalities. We first use three modality-specific encoders $En^m(\cdot)$ to project $\boldsymbol{U}_o^m$ into latent vectors $\boldsymbol{Z}_o^m \in \mathbb{R}^{\mathcal{C} \times d_z}$ in the latent space $\mathcal{S}^z$. From the space $\mathcal{S}^z$, we calculate the intrinsic vectors $\boldsymbol{D}^m = \{\boldsymbol{d}_j^m \in \mathbb{R}^{d_z}\}_{j=1}^{\mathcal{C}}$ to reflect the feature distribution of each label j in different modalities (explained in the next sub-section). Then, three modality-specific decoders $De^m(\cdot)$ transform vectors $\boldsymbol{Z}_o^m$ and $\boldsymbol{D}^m$ back to decoded vectors $\tilde{\boldsymbol{U}}_o^m, \tilde{\boldsymbol{D}}^m \in \mathbb{R}^{\mathcal{C} \times d}$ respectively.

To realize cross-modal feature fusion, we employ a two-level reconstruction process with three networks $f^{va2t}(\cdot)$, $f^{ta2v}(\cdot)$ and $f^{tv2a}(\cdot)$ (detailed analysis in Appendix A). Taking the modality t as an example, we first concatenate the intrinsic features $\tilde{\boldsymbol{D}}^t$ and semantic features $\tilde{\boldsymbol{U}}_o^{\{v,a\}}$ in a certain modality order, where the former reflects the feature distribution of the current modality (t) and the latter provides semantic information of other modalities (v, a). The concatenated vectors are input into $f^{va2t}(\cdot)$ to obtain the first-level reconstruction representations (FRR) $\boldsymbol{U}_\alpha^t \in \mathbb{R}^{\mathcal{C} \times d}$. Then, $\boldsymbol{U}_\alpha^m$ of all modalities are concatenated and feed into $f^{va2t}(\cdot)$ to generate the second-level reconstruction representations (SRR) $\boldsymbol{U}_\beta^t \in \mathbb{R}^{\mathcal{C} \times d}$. The reconstruction-based fusion process of all modalities is expressed as,

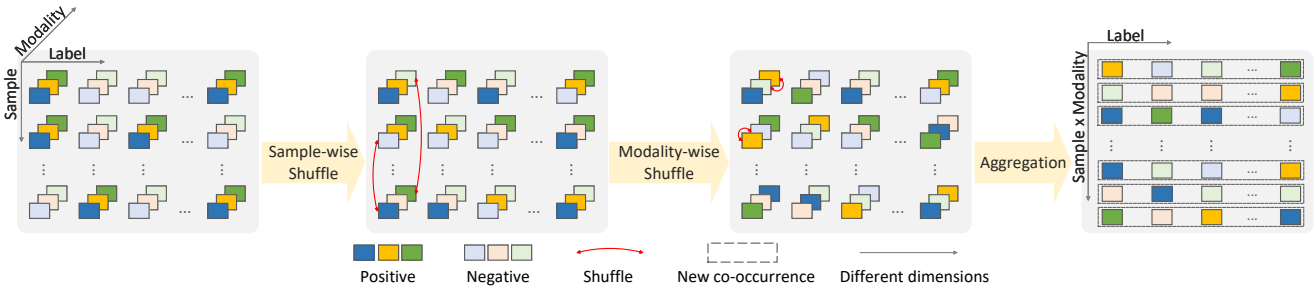

Figure 3: The semantic diagram of shuffle. The shuffle is conducted in both batch and modality dimensions. Different colors represent different modalities. After shuffling, new co-occurrence relations can appear in the training set (the gray dotted boxes).

$$\begin{aligned}
\boldsymbol{U}_\alpha^t &= f^{va2t}([\tilde{\boldsymbol{D}}^t;\tilde{\boldsymbol{U}}_o^v;\tilde{\boldsymbol{U}}_o^a]), \boldsymbol{U}_\beta^t = f^{va2t}([\boldsymbol{U}_\alpha^t;\boldsymbol{U}_\alpha^v;\boldsymbol{U}_\alpha^a]),\\
\boldsymbol{U}_\alpha^v &= f^{ta2v}([\tilde{\boldsymbol{U}}_o^t;\tilde{\boldsymbol{D}}^v;\tilde{\boldsymbol{U}}_o^a]), \boldsymbol{U}_\beta^v = f^{ta2v}([\boldsymbol{U}_\alpha^t;\boldsymbol{U}_\alpha^v;\boldsymbol{U}_\alpha^a]),\\
\boldsymbol{U}_\alpha^a &= f^{tv2a}([\tilde{\boldsymbol{U}}_o^t;\tilde{\boldsymbol{U}}_o^v;\tilde{\boldsymbol{D}}^a]), \boldsymbol{U}_\beta^a = f^{tv2a}([\boldsymbol{U}_\alpha^t;\boldsymbol{U}_\alpha^v;\boldsymbol{U}_\alpha^a]).
\end{aligned} \tag{2}$$

To ensure that the reconstructed feature vectors can restore the original information, we use the mean square error to formulate the reconstruction loss as:

$$\mathcal{L}_{rec} = \sum_m^M \left(\|\boldsymbol{U}_o^m - \tilde{\boldsymbol{U}}_o^m\|_F + \|\boldsymbol{U}_o^m - \boldsymbol{U}_\alpha^m\|_F\right), \tag{3}$$

where $\|\cdot\|_F$ returns the Frobenius norm of the matrix.

Due to the modality heterogeneity, different modalities express each emotion with different contributions. Therefore, we introduce a Max Pooling-like network to impel each label to focus on its most relevant modality. Specifically, we utilize three modality-specific classifiers $h_{\{t,v,a\}}(\cdot)$ on $\boldsymbol{U}_o^m, \boldsymbol{U}_\alpha^m, \boldsymbol{U}_\beta^m$ to calculate label prediction for $\{t,v,a\}$ modalities, respectively. Then, we connect a Max-Pooling layer on these predictions to filter the most relevant modality of each label. Taking $\boldsymbol{U}_o^m$ as an example, the final output via the above network is calculated as,

$$\boldsymbol{s}^o = \text{MaxPool}\big(h_t(\boldsymbol{U}_o^t), h_v(\boldsymbol{U}_o^v), h_a(\boldsymbol{U}_o^a)\big) \in \mathbb{R}^{\mathcal{C}}. \tag{4}$$

In the same way, we can also obtain $\boldsymbol{s}^\alpha$ and $\boldsymbol{s}^\beta$. Finally, we calculate the binary cross entropy (BCE) losses as,

$$\mathcal{L}_{cls}^{lsr} = \gamma_o l(\boldsymbol{s}^o, \boldsymbol{y}) + \gamma_\alpha l(\boldsymbol{s}^\alpha, \boldsymbol{y}) + \gamma_\beta l(\boldsymbol{s}^\beta, \boldsymbol{y}), \tag{5}$$

where l is the BCE loss and $\gamma_{o,\alpha,\beta}$ are trade-off parameters.

Contrastive Representation Learning. To enable intrinsic vectors $\boldsymbol{D}^m$ to reflect the feature distribution of each label in different modalities, we utilize contrastive learning to learn a distinguishable latent embedding space $\mathcal{S}^z$. For samples in a batch of size B, after obtaining $\boldsymbol{U}_o^m$, $\boldsymbol{U}_\alpha^m$, $\boldsymbol{U}_\beta^m$, we feed them into the corresponding encoder $En^m(\cdot)$ to generate L_2-normalized latent embeddings $\boldsymbol{Z}_o^m$, $\boldsymbol{Z}_\alpha^m$, $\boldsymbol{Z}_\beta^m \in \mathbb{R}^{\mathcal{C}\times d_z}$, respectively. We follow the SCL (Khosla et al. 2020) and additionally maintain a queue storing the most current latent embeddings, and we update the queue chronologically. Thus, we have the contrastive embedding pool as $\boldsymbol{E} = \{\boldsymbol{Z}_{\{o,\alpha,\beta\}}^{\{t,v,a\}}\}_{i=1}^B \cup$ queue. Given an anchor embedding $\boldsymbol{e} \in \boldsymbol{E}$, the contrastive loss is defined by contrasting its positive set with the remainder of the pool $\boldsymbol{E}$ as,

$$\mathcal{L}_{scl}(\boldsymbol{e}, \boldsymbol{E}) = -\frac{1}{|\boldsymbol{P}(\boldsymbol{e})|} \sum_{\boldsymbol{e}^+ \in \boldsymbol{P}(\boldsymbol{e})} \log \frac{\exp\left(\boldsymbol{e}^\top \boldsymbol{e}^+/\tau\right)}{\sum\limits_{\boldsymbol{e}' \in \boldsymbol{E}(\boldsymbol{e})} \exp\left(\boldsymbol{e}^\top \boldsymbol{e}'/\tau\right)}, \tag{6}$$

where $\boldsymbol{P}(\boldsymbol{e})$ is the positive set and $\boldsymbol{E}(\boldsymbol{e}) = \boldsymbol{E}\backslash\{\boldsymbol{e}\}$. $\tau \in \mathbb{R}^+$ is the temperature. The contrastive loss of the batch is:

$$\mathcal{L}_{scl} = \sum\nolimits_{\boldsymbol{e}\in\boldsymbol{E}} \mathcal{L}_{scl}(\boldsymbol{e}, \boldsymbol{E}). \tag{7}$$

To construct the positive set, considering the purpose of learning the modality-specific feature distribution of each label, we redefine the label for each $\boldsymbol{e}$. According to the modality m, label category j and label polarity k, the new label is defined as $\tilde{y} = l_{j,k}^m, m \in \{t,v,a\}, j \in [\mathcal{C}], k \in \{pos, neg\}$. Thus, the positive examples are selected as $\boldsymbol{P}(\boldsymbol{e}) = \{\boldsymbol{e}' | \boldsymbol{e}' \in \boldsymbol{E}(\boldsymbol{e}), \tilde{y}' = \tilde{y}\}$, where $\tilde{y}'$ is the label for $\boldsymbol{e}'$. In other words, the positive set is those embeddings from the same modality with the same label category and polarity. Importantly, we keep a prototype embedding $\boldsymbol{\mu}_{j,k}^m \in \mathbb{R}^{d_z}$ corresponding to each class $l_{j,k}^m$, which can be deemed as a set of representative embedding vectors. To reduce the computational toll and training latency, we update the class-conditional prototype vector in a moving-average style as,

$$\boldsymbol{\mu}_{j,k}^m = \text{Normalize}(\phi\boldsymbol{\mu}_{j,k}^m + (1-\phi)\boldsymbol{e}), \quad \text{if } \tilde{y} = l_{j,k}^m, \tag{8}$$

where the momentum prototype $\boldsymbol{\mu}_{j,k}^m$ is defined by the moving average of the normalized embedding whose defined class conforms to $l_{j,k}^m$. ϕ is a hyperparameter. During training, we leverage prototypes to obtain the intrinsic vectors $\boldsymbol{D}^m = [\boldsymbol{d}_1^m, \ldots, \boldsymbol{d}_{\mathcal{C}}^m]$ via the soft-max pattern as,

$$\boldsymbol{d}_j^m = \sum_k^{\{pos,neg\}} \alpha_{j,k}^m \boldsymbol{u}_{j,k}^m, \ \alpha_{j,k}^m = \frac{\exp(\boldsymbol{e}^\top \boldsymbol{u}_{j,k}^m)}{\sum\limits_{k'}^{\{pos,neg\}} \exp(\boldsymbol{e}^\top \boldsymbol{u}_{j,k'}^m)}, \tag{9}$$

while during prediction, the hard-max pattern is used as,

$$\boldsymbol{d}_j^m = I_{[\alpha_{j,pos}^m > \alpha_{j,neg}^m]} \boldsymbol{u}_{j,pos}^m + I_{[\alpha_{j,pos}^m \leq \alpha_{j,neg}^m]} \boldsymbol{u}_{j,neg}^m, \tag{10}$$

where $I_{[\cdot]}$ is the indicator function.

Shuffle-based Feature Aggregation

Although exploiting the most relevant modality is sufficient to find discriminative features, multi-modal fusion can use complementary information to obtain more robust representations. Therefore, we design a shuffle-based aggregation to exploit cross-modal information, which includes sample- and modality-wise shuffle processes. The motivations of the sample- and modality-wise shuffle are to enrich the co-occurrence relations of labels and realize random cross-modal aggregation, respectively. As shown in Figure 3, after obtaining the SRR of a batch of samples, two shuffle processes are performed sequentially and independently. Specifically, we stack the vectors $\boldsymbol{U}_\beta^m$ of the batch as $\boldsymbol{V} = \text{Stack}\big(\big\{[\boldsymbol{U}_\beta^t; \boldsymbol{U}_\beta^v; \boldsymbol{U}_\beta^a]_i\big\}_{i=1}^B\big) \in \mathbb{R}^{B\times M\times \mathcal{C}\times d}$, where M is the number of modalities. Firstly, on each modality m, we perform the sample-wise shuffle (sws) as,

$$\boldsymbol{V}_{[:,m]} := [\boldsymbol{v}_1^m, \ldots, \boldsymbol{v}_B^m] \overset{\text{sws}}{\rightarrow} \tilde{\boldsymbol{V}}_{[:,m]} := \left[\boldsymbol{v}_{r_1}^m, \ldots, \boldsymbol{v}_{r_B}^m\right], \tag{11}$$

where $\{r_i\}_1^B$ are new indices of samples. Then, for each sample, the modality-wise shuffle (mws) is performed as,

$$\tilde{\boldsymbol{V}}_{[i,:]} := \left[\tilde{\boldsymbol{v}}_i^1, \ldots, \tilde{\boldsymbol{v}}_i^M\right] \overset{\text{mws}}{\rightarrow} \hat{\boldsymbol{V}}_{[i,:]} := [\tilde{\boldsymbol{v}}_i^{r_1}, \ldots, \tilde{\boldsymbol{v}}_i^{r_M}], \tag{12}$$

where $\{r_i\}_1^M$ are new indices of modalities. Then, $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$ are concatenated on the label dimension, respectively, as,

$$\begin{aligned}\boldsymbol{Q} &= \left\{\left\{\boldsymbol{q}_i^m = [\boldsymbol{v}_{i,1}^m; \ldots; \boldsymbol{v}_{i,\mathcal{C}}^m]\right\}_m^{\{t,v,a\}}\right\}_i^B \in \mathbb{R}^{B\times M\times \mathcal{C}\cdot d},\\ \hat{\boldsymbol{Q}} &= \left\{\left\{\hat{\boldsymbol{q}}_i^m = [\hat{\boldsymbol{v}}_{i,1}^m; \ldots; \hat{\boldsymbol{v}}_{i,\mathcal{C}}^m]\right\}_m^{\{t,v,a\}}\right\}_i^B \in \mathbb{R}^{B\times M\times \mathcal{C}\cdot d}.\end{aligned} \tag{13}$$

It is worth noting that unlike $\boldsymbol{q}_i^m$, which is concatenated by features from a single modality and a single sample, the features constituting $\hat{\boldsymbol{q}}_i^m$ are randomly sampled from 1 to M modalities and 1 to $\mathcal{C}$ samples. Finally, the $\boldsymbol{Q}$ and $\hat{\boldsymbol{Q}}$ are used to fine-tune a classifier $h_c(\cdot)$ with the BCE loss as,

$$\mathcal{L}_{cls}^{agg} = \frac{1}{M}\sum_m^M \big(l(h_c(\boldsymbol{q}^m), \boldsymbol{y}_{\boldsymbol{q}^m}) + \gamma_{sf} l(h_c(\hat{\boldsymbol{q}}^m), \boldsymbol{y}_{\hat{\boldsymbol{q}}^m})\big), \tag{14}$$

where γ_{sf} is the trade-off parameter. Combing the Equation 3, 5, 7 and 14, the final objective function is formulated as,

$$\mathcal{L} = \mathcal{L}_{cls}^{agg} + \mathcal{L}_{cls}^{lsr} + + \gamma_s \mathcal{L}_{scl} + \gamma_r \mathcal{L}_{rec}, \tag{15}$$

where γ_s, γ_r are trade-off parameters. During prediction, to utilize both the most relevant modality and multi-modal fusion, the prediction of the test sample i' is obtained as,

$$\hat{\boldsymbol{y}}_{i'} = \frac{1}{2}\big(\frac{1}{M}\sum\nolimits_m^M h_c(\boldsymbol{q}_{i'}^m) + \boldsymbol{s}_{i'}^\beta\big) \in \mathbb{R}^{\mathcal{C}}. \tag{16}$$

Experiments

Experimental Settings

Dataset and Evaluation Metrics. We evaluate CARAT on two benchmark MMER datasets (CMU-MOSEI (Zadeh et al. 2018b) and M^{3}ED (Zhao et al. 2022)), which maintained settings in the public SDK[2] . Four evaluation metrics are employed: Accuracy (Acc), Micro-F1, Precision (P), and Recall (R). More detailed descriptions and preprocessing of datasets are shown in Appendix B.

Baselines. We compare CARAT with various approaches of two groups. The first group is Multi-Label Classification (MLC) methods. Specifically, in these approaches, the multi-modal inputs are early fused (simply concatenated) as a new input. For classic methods: (1) **BR** (Boutell et al. 2004) transforms MLC into multiple binary classifications while ignoring label correlations. (2) **LP** (Tsoumakas and Katakis 2006) breaks the initial label set into several random subsets and trains a corresponding classifier. (3) **CC** (Read et al. 2011) transforms MLC into a chain of binary classification problems by considering high-order label correlations. For deep-based methods: (4) **SGM** (Yang et al. 2018) views MLC as a sequence generation problem via label correlation. (5) **LSAN** (Xiao et al. 2019) explores the semantic connection between labels and documents to construct label-specific document representation. (6) **ML-GCN** (Chen et al. 2019b) employs GCN to map label representations and captures label correlations for image recognition.

The second group is multi-modal multi-label methods. (7) **MulT** (Tsai et al. 2019) uses cross-modal interactions to fuse information from one modality to another. (8) **MISA** (Hazarika, Zimmermann, and Poria 2020) learns modality-invariant and modality-specific representations for the fusion. (9) **MMS2S** (Zhang et al. 2020) handles the modality and label dependence in a sequence-to-set approach. (10) **HHMPN** (Zhang et al. 2021) models feature-to-label, label-to-label and modality-to-label dependencies via graph message passing. (11) **TAILOR** (Zhang et al. 2022) adversarially depicts commonality and diversity among modalities to obtain discriminative representations. (12) **AMP** (Ge et al. 2023) learns robust representations with adversarial temporal masking and Adversarial Parameter Perturbation.

Implementation Details. We set the size of hidden states as $d = 256, d_z = 64$. The size of the embedding queue is set to 8192. All encoders $En^m(\cdot)$ and decoders $De^m(\cdot)$ are implemented by 2-layer MLPs. We set hyper-parameters $\gamma_o = 0.01$, $\gamma_\alpha = 0.1$, $\gamma_\beta = 1$, $\gamma_s = 1$, $\gamma_{sf} = 0.1$ and $\gamma_r = 1$ and the analysis of different weight settings is presented in Appendix A. We set $l_t = 6, l_v = l_a = 4$ for the layer number of Transformer Encoders. We employ the Adam (Kingma and Ba 2014) optimizer with the initial learning rate of $5e^{-5}$ and a liner decay learning rate schedule with a warm-up strategy. The batch size B is set to 64. During training, we train methods for 20 epochs to select the model with the best F1 score on the validation set as our final model. All experiments are conducted with one NVIDIA A100 GPU.

Experimental Results

Performance Comparison. We show performance comparisons on CMU-MOSEI and M^{3}ED (only partial multi-

[2]https://github.com/A2Zadeh/CMU-MultimodalSDK

[3]https://github.com/AIM3-RUC/RUCM3ED

Approaches	Methods	Aligned				Unaligned			
		Acc	P	R	Micro-F1	Acc	P	R	Micro-F1
Classical	BR	0.222	0.309	0.515	0.386	0.233	0.321	0.545	0.404
	LP	0.159	0.231	0.377	0.286	0.185	0.252	0.427	0.317
	CC	0.225	0.306	0.523	0.386	0.235	0.320	0.550	0.404
Deep-based	SGM	0.455	0.595	0.467	0.523	0.449	0.584	0.476	0.524
	LSAN	0.393	0.550	0.459	0.501	0.403	0.582	0.460	0.514
	ML-GCN	0.411	0.546	0.476	0.509	0.437	0.573	0.482	0.524
Multi-modal	MulT	0.445	0.619	0.465	0.531	0.423	0.636	0.445	0.523
	MISA	0.430	0.453	**0.582**	0.509	0.398	0.371	**0.571**	0.450
	MMS2S	0.475	0.629	0.504	0.560	0.447	0.619	0.462	0.529
	HHMPN	0.459	0.602	0.496	0.556	0.434	0.591	0.476	0.528
	TAILOR	0.488	0.641	0.512	0.569	0.460	0.639	0.452	0.529
	AMP	0.484	0.643	0.511	0.569	0.462	0.642	0.459	0.535
	CARAT	**0.494**	**0.661**	0.518	**0.581**	**0.466**	**0.652**	0.466	**0.544**

Table 1: Performance comparison between CARAT and baselines on CMU-MOSEI dataset with aligned and unaligned settings.

Methods	Acc	P	R	Micro-F1
MMS2S	0.645	0.813	0.737	0.773
HHMPN	0.648	0.816	0.743	0.778
TAILOR	0.647	0.814	0.739	0.775
AMP	0.654	0.819	0.748	0.782
CARAT	**0.664**	**0.824**	**0.755**	**0.788**

Table 2: Performance comparison on the M^{3}ED dataset.

modal baselines) in Table 1,2, and observations are as follows: 1) CARAT significantly outperforms all rivals by a significant margin. Although **MISA** has a prominent recall, its precision drops to a poor value and its performance is far inferior to CARAT on more important metrics (Micro-F1 and accuracy). Furthermore, CARAT still maintains a decent performance boost in the unaligned setting, which proves that CARAT can break the barrier of the modality gap better than others. 2) Among uni-modal approaches, the superior performance of deep-based methods, i.e. **SGM**, **LSAN** and **ML-GCN**, over classic methods, i.e. **BR**, **CC** and **LP**, indicates that deep representation can better capture semantic features and label correlations help to capture more meaningful information. 3) Compared with uni-modal approaches, multi-modal methods typically exhibit better performance, which shows the necessity of modeling multi-modal interactions. 4) Among all baselines, **TAILOR** achieves competitive performance, which validates the effectiveness of leveraging commonality and diversity among modalities to obtain the discriminative label representations.

Ablation Study. To demonstrate the importance of each component, we compare CARAT with various ablated variants. As shown in Table 3, we can see:

1) *Effect of exploiting both specificity and complementarity:* By using both features of the most relevant modality (MRM) and aggregated features (AGG), (1) is better than (2) and (3), which indicates the significance of binding modality specificity and complementarity.

2) *Effect of the contrastive representation learning:* Without conducting the loss $\mathcal{L}_{scl}$, (4) is worse than CARAT, which illustrates the significance of leveraging contrastive learning to learn distinguishable representations. Further, by removing the process of encoding and decoding, (5) is worse than (4), which validates the rationality of exploring the intrinsic embeddings in the latent space.

Approaches	Acc	P	R	Micro-F1
(1) MRM + AGG	0.475	0.647	0.507	0.569
(2) only MRM	0.474	0.641	0.502	0.563
(3) only AGG	0.472	0.639	0.506	0.565
(4) w/o $\mathcal{L}_{scl}$	0.481	0.640	0.515	0.571
(5) w/o En, De	0.475	0.638	0.514	0.569
(6) w/o $\mathcal{L}_{rec}$	0.482	0.644	0.516	0.573
(7) w/o α-recon	0.483	0.636	**0.520**	0.572
(8) w/o β-recon	0.482	0.631	0.513	0.566
(9) w/o $\alpha\&\beta$-recon	0.475	0.619	0.503	0.555
(10) w/o sw-shf	0.491	0.659	0.511	0.575
(11) w/o mw-shf	0.490	0.656	0.514	0.576
(12) w/o shf	0.489	0.658	0.509	0.574
(13) CARAT	**0.494**	**0.661**	0.518	**0.581**

Table 3: Ablation tests on the aligned CMU-MOSEI. "MRM" and "AGG" respectively denote using features of the Most Relevant Modality and AGGregated features. "w/o" means removing. "w/o En, De" denotes removing the encoding and decoding. "w/o $\{\alpha, \beta, \alpha\&\beta\}$-recon" denotes removing the first-level reconstruction, second-level reconstruction, or both. "w/o { sw-, mw-, ϕ}shf" denotes removing the sample-, modality-wise shuffle process or both. Detailed implementations are in Appendix A.

3) *Effect of the two-level feature reconstruction:* First, (6) is worse than CARAT, which reveals the effectiveness of using loss $\mathcal{L}_{rec}$ to constrain feature reconstruction. Removing the first- and second-level reconstruction processes, (7) and (8) have different degrees of performance degradation compared to CARAT. When the entire reconstruction process is removed, the performance of (9) is further reduced than (7) and (8), which confirms the effectiveness of multi-level feature reconstruction to achieve multi-modal fusion.

4) *Effect of different shuffling operations:* Excluding any round of the shuffling process, (10) and (11) are worse than CARAT, and (12) is even worse when both shuffling pro-

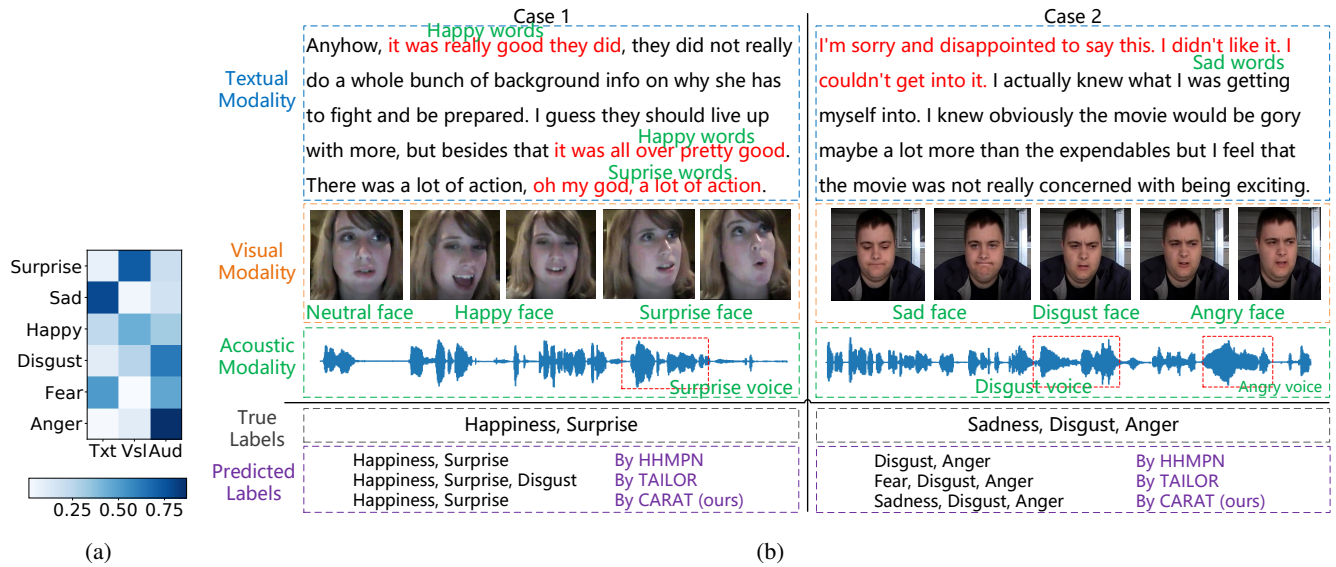

Figure 4: (a) The visualization of modality-to-label dependencies, indicating the correlation of labels in each row to modality in each column, where darker colors indicate stronger correlations. (b) Two cases of emotion recognition by multiple methods.

cesses are removed, which confirms the effectiveness of performing shuffling in both sample and modality dimensions.

Figure 5: t-SNE visualization of embeddings without/with (left/right) CL. Different colors represent different modalities and different shades represent different emotions.

Analysis

Visualization of Distinguishable Representations. To investigate the efficacy of contrastive learning (CL) on distinguishable representations, we visualize embeddings $\boldsymbol{Z}_o^m$ in space $\mathcal{S}^z$ using t-SNE (Van der Maaten and Hinton 2008) without or with contrastive learning on the aligned CMU-MOSEI dataset. As shown in Figure 5, without CL (left subfigure), although embeddings belonging to different modalities can be well distinguished, the embeddings of different classes in each modality are lumped together. In contrast, in the right subfigure, embeddings belonging to different modalities and labels are explicitly separated from each other, and embeddings of the same modality still maintain a tighter distribution. Thus, with CL, CARAT produces well-separated clusters and more distinguishable representations.

Visualization of Modality-to-label Correlations. We visualize the correlation of labels with their most relevant modalities. As shown in Figure 4 (a), each label focuses on different modalities unequally and usually has its own prone modality. The modality-to-label correlations differ from label to label, e.g., emotion *surprise* and *sad* are highly correlated with the visual and textual modality, respectively. More visualization examples are shown in Appendix E.

Case Study. To further demonstrate the effectiveness of CARAT, Figure 4 (b) presents two cases. 1) Emotions expressed by different modalities are not consistent, which reflects the modality specificity. E.g., in Case 2, emotion *angry* can be mined intuitively from the visual and audio, but not the text. 2) HHMPN wrongly omitted relevant labels due to neglecting modality specificity, which results in the inability to capture richer semantic information. In contrast, TAILOR gives wrong related labels. Since TAILOR uses self-attention that can only explore label correlations within each sample, global information cannot be exploited. Overall, our CARAT achieves the best performance.

Conclusion

In this paper, we propose ContrAstive feature Reconstruction and AggregaTion (CARAT) for MMER, which integrates effective representation learning and multiple dependency modeling into a unified framework. We propose a reconstruction-based fusion mechanism by contrastively learning modal-separated and label-specific features to model fine-grained modality-to-label dependencies. To further exploit the modality complementarity, we introduce a shuffle-based aggregation strategy to enrich co-occurrence collaboration among labels. Experiments on benchmark datasets CMU-MOSEI and M^{3}ED demonstrate the effectiveness of CARAT over state-of-the-art methods.

Acknowledgments

This work is supported by the National Key R&D Program of China (No.2022YFB3304100) and the Pioneer R&D Program of Zhejiang "Data Preparation, Optimization, and Augmentation for Domain-Specific Large Models".

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal machine learning: A survey and taxonomy. *Trans. Pattern Anal. Mach. Intell.*, 41(2): 423–443.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern recognition*, 37(9): 1757–1771.

Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019a. Learning semantic-specific graph representation for multi-label image recognition. In *ICCV*, 522–531.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019b. Multi-label image recognition with graph convolutional networks. In *CVPR*, 5177–5186.

Franceschini, R.; Fini, E.; Beyan, C.; Conti, A.; Arrigoni, F.; and Ricci, E. 2022. Multimodal Emotion Recognition with Modality-Pairwise Unsupervised Contrastive Loss. In *ICPR*, 2589–2596.

Galik, M.; and Rank, S. 2012. Modelling emotional trajectories of individuals in an online chat. In *MATES*, volume 7598, 96–105.

Ge, S.; Jiang, Z.; Cheng, Z.; Wang, C.; Yin, Y.; and Gu, Q. 2023. Learning Robust Multi-Modal Representation for Multi-Label Emotion Recognition via Adversarial Masking and Perturbation. In *WWW*, 1510–1518.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP*, 154–164.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *MM*, 1122–1131.

Hazirbas, C.; Ma, L.; Domokos, C.; and Cremers, D. 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *ACCV*, volume 10111, 213–228.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9726–9735.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Huang, J.; Li, G.; Huang, Q.; and Wu, X. 2016. Learning label-specific features and class-dependent labels for multi-label classification. *Trans. Knowl. Data Eng.*, 28(12): 3309–3323.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Li, J.; Zhou, P.; Xiong, C.; and Hoi, S. C. 2021. Prototypical contrastive learning of unsupervised representations. In *ICLR*.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2247–2256.

Ma, Q.; Yuan, C.; Zhou, W.; and Hu, S. 2021. Label-specific dual graph neural network for multi-label text classification. In *ACL*, 3855–3864.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *ICML*, 689–696.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, 6892–6899.

Pham, H.; Manzini, T.; Liang, P. P.; and Poczos, B. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. *arXiv preprint arXiv:1807.03915*.

Ramachandram, D.; and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *Signal Process. Mag.*, 34(6): 96–108.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning*, 5782: 254–269.

Tsai, C.-P.; and Lee, H.-Y. 2020. Order-free learning alleviating exposure bias in multi-label classification. In *AAAI*, 6038–6045.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 6558–6569.

Tsoumakas, G.; and Katakis, I. 2006. Multi-label classification: An overview international journal of data warehousing and mining. *The label powerset algorithm is called PT3*, 3(3).

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Xiao, L.; Huang, X.; Chen, B.; and Jing, L. 2019. Label-specific document representation for multi-label text classification. In *EMNLP*, 466–475.

Yang, P.; Luo, F.; Ma, S.; Lin, J.; and Sun, X. 2019. A deep reinforced sequence-to-set model for multi-label classification. In *ACL*, 5252–5258.

Yang, P.; Sun, X.; Li, W.; Ma, S.; Wu, W.; and Wang, H. 2018. SGM: sequence generation model for multi-label classification. In *COLING*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 1103–1114.

Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018a. Multi-attention recurrent network for human communication comprehension. In *AAAI*, 5642–5649.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, 2236–2246.

Zhang, D.; Ju, X.; Li, J.; Li, S.; Zhu, Q.; and Zhou, G. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *EMNLP*, 3584–3593.

Zhang, D.; Ju, X.; Zhang, W.; Li, J.; Li, S.; Zhu, Q.; and Zhou, G. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *AAAI*, 14338–14346.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *Trans. Knowl. Data Eng.*, 26(8): 1819–1837.

Zhang, Y.; Chen, M.; Shen, J.; and Wang, C. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *AAAI*, 9100–9108.

Zhao, J.; Zhang, T.; Hu, J.; Liu, Y.; Jin, Q.; Wang, X.; and Li, H. 2022. M3ED: Multi-modal multi-scene multi-label emotional dialogue database. In *ACL*.

Zheng, Z.; An, G.; Cao, S.; Yang, Z.; and Ruan, Q. 2022. PromptLearner-CLIP: Contrastive Multi-Modal Action Representation Learning with Context Optimization. In *ACCV*, volume 13844, 554–570.

Zhu, S.; Li, S.; and Zhou, G. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *ACL*, 471–480.

Zolfaghari, M.; Zhu, Y.; Gehler, P.; and Brox, T. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *ICCV*, 1430–1439.