Tailor Versatile Multi-modal Learning for Multi-label Emotion Recognition

Yi Zhang, Mingyuan Chen, Jundong Shen, Chongjun Wang[*]

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{njuzhangy, mychen, jdshen}@smail.nju.edu.cn, {chjwang}@nju.edu.cn

Abstract

Multi-modal Multi-label Emotion Recognition (MMER) aims to identify various human emotions from heterogeneous visual, audio and text modalities. Previous methods mainly focus on projecting multiple modalities into a common latent space and learning an identical representation for all labels, which neglects the diversity of each modality and fails to capture richer semantic information for each label from different perspectives. Besides, associated relationships of modalities and labels have not been fully exploited. In this paper, we propose versaTile multi-modAl learning for multI-labeL emOtion Recognition (TAILOR), aiming to refine multi-modal representations and enhance discriminative capacity of each label. Specifically, we design an adversarial multi-modal refinement module to sufficiently explore the commonality among different modalities and strengthen the diversity of each modality. To further exploit label-modal dependence, we devise a BERT-like cross-modal encoder to gradually fuse private and common modality representations in a granularity descent way, as well as a label-guided decoder to adaptively generate a tailored representation for each label with the guidance of label semantics. In addition, we conduct experiments on the benchmark MMER dataset CMU-MOSEI in both aligned and unaligned settings, which demonstrate the superiority of TAILOR over the state-of-the-arts. Code is available at https://github.com/kniter1/TAILOR.

Introduction

In real-world applications, videos are often characterized by heterogeneous representations (i.e., visual, audio and text) and annotated with various emotion labels (e.g., *happy*, *surprise*). Multi-modal Multi-label Emotion Recognition (MMER) (Ju et al. 2020; Zhang et al. 2021a) refers to identifying various emotions by leveraging visual, audio and text modalities presented in videos.

Multi-modal learning (Baltrusaitis, Ahuja, and Morency 2019) processes heterogeneous information collected from multiple sources, which gives rise to two emergent issues: intra-modal representation and inter-modal fusion. Intra-modal representation learning mainly exploits consistency and complementarity of multiple modalities to bridge the heterogeneous modality gap. Previous methods project each
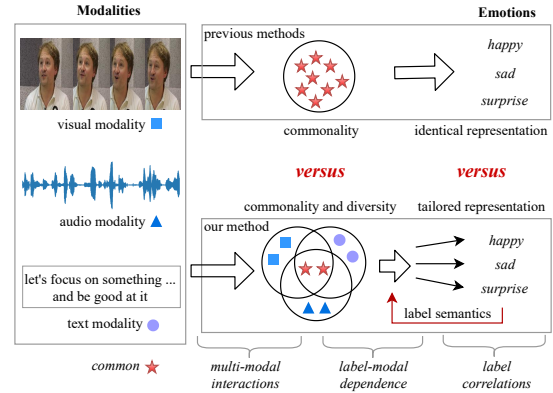
---

[*]Corresponding author

Figure 1: Previous methods versus our method.

modality into a shared latent space to eliminate redundancy. However, they neglect the fact that different modalities reveal distinctive characteristic of emotions from different perspectives. Concerning the fusion manner, existing inter-modal fusion methods can be divided into: aggregation-based fusion, alignment-based fusion and the mixture of them (Baltrusaitis, Ahuja, and Morency 2019). Aggregation-based fusion adopts concatenation (Ngiam et al. 2011), tensor (Zadeh et al. 2017) or attention (Zadeh et al. 2018b) to combine multiple modalities. Alignment-based fusion centers on latent cross-modal adaptation, which adapts streams from one modality to another (Tsai et al. 2019). The key challenge of multi-modal learning lies in 1) how to integrate commonality while preserving diversity of each individual modality; 2) how to align different modality distributions interactively for inter-modal fusion.

Multi-label learning (Zhang and Zhou 2014) deals with rich semantic meanings of complicated objects, where label correlations are considered as the key to effective multi-label learning (Zhu, Kwok, and Zhou 2018). Many methods exploit label correlations by the similarity between label vectors, and then seamlessly incorporated into an identical representation. However, they are unable to reflect collaboration relationships among labels. On the other hand, many researches have been developed to improve the performance by learning label-specific representations in a label-by-label manner, which are generated independently and may lead to suboptimal problem due to the ignorance of label corre-

lations (Zhang, Fang, and Wang 2021). The key challenge of multi-label learning is how to effectively encode inherent and discriminative characteristics of each label in both the feature space and the label space.

To address the above challenges, we propose versaTile multi-modAl learning for multI-labeL emOtion Recognition (TAILOR), which sufficiently copes with modality heterogeneity and label heterogeneity. To bridge the heterogeneity gap, we capture modality interactions, label correlations and label-modal dependence in the following 3 spaces.

1) *In the modality feature space*, we emphasis less on pretraining. For intra-modal representation, we devise an adversarial network to explicitly extract commonality and diversity, constrained by common semantics and orthogonality. For inter-modal fusion, we propose a novel granularity-based fusion with BERT-like transformer encoder.

2) *In the label space*, we adopt self-attention (Vaswani et al. 2017) to exploit high-order label correlations, which can be further integrated to capture label semantics.

3) *To bridge the gap between modality feature space and label space*, we adapt Transformer decoder to align fused multi-modal representations with label semantics, which aims to learn tailored representation of each label with the guidance of label semantics.

Figure 1 illustrates the difference between previous methods and our proposed method. The main contributions can be summarized as follows.

- A novel framework of versaTile multi-modAl learning for multI-labeL emOtion Recognition (TAILOR) is proposed, which adversarially depicts commonality and diversity among multiple modalities, as well as enhance discriminative capability of label representations.

- TAILOR adversarially extracts private and common modality representations. Then a BERT-like transformer encoder is devised to gradually fuse these representations in a granularity descent way, which is incorporated with label semantics to generate tailored label representation.

- Extensive experiments conducted on benchmark CMU-MOSEI dataset demonstrate the excellent performance of TAILOR in both aligned and unaligned settings.

Related Work

Emotion recognition is broadly studied with **uni-modal** (Yang et al. 2018; Majumder et al. 2019; Saha et al. 2020; Jiao, Lyu, and King 2020; Huang et al. 2021), **bi-modal** (Mittal et al. 2020b; Liu et al. 2020; Zhao et al. 2020) and **multi-modal** (Mittal et al. 2020a; Sun et al. 2020; Zhang et al. 2021a; Lv et al. 2021). More effective **multi-modal fusion** translates to better performance. The most straightforward way is to directly concatenating feature maps from each modality (Ngiam et al. 2011). To leverage complementary information across different modalities, tensor fusion (Zadeh et al. 2017; Liu et al. 2018), memory fusion (Zadeh et al. 2018a), factorization fusion (Valada, Mohan, and Burgard 2020) explicitly account for intra-modal and inter-modal dynamics. The above mentioned methods are aggregation-based fusion and the modality gap heavily affects cross-modal fusion. To bridge the modality gap,

GAN (Goodfellow et al. 2014) has attracted significant interest in learning joint distributions between bi-modal or multi-modal (Pham et al. 2018; Tsai et al. 2018; Pham et al. 2019; Mai, Hu, and Xing 2020), alignment-based fusion (Baltrusaitis, Ahuja, and Morency 2019) latently adapts streams from one modality to another via Transformer (Goodfellow et al. 2014). Even though, they tend to fuse into a joint embedding space, which neglects the specificity of each modality. (Wang et al. 2020b) adapts the fusion of modality-specific streams and fuses only the relevant complementary information. For example, (Wu et al. 2019; Hazarika et al. 2020) integrates the common information across modalities, meanwhile preserving the specific patterns of each modality.

In multi-label learning, modeling **label correlations** has been proven to be an effective strategy (Zhang and Zhou 2014; Zhu, Kwok, and Zhou 2018; Feng, An, and He 2019; Wang et al. 2020a). It might be suboptimal to learn a subset of features shared by all the labels. Another significant strategy is **label-specific learning** (Zhang and Wu 2014; Huang et al. 2016; Zhang, Fang, and Wang 2021), where each label is determined by some discriminative characteristics, e.g., visual attention (Chen et al. 2019a,b) and textual attention (Xiao et al. 2019; Zhang et al. 2021b).

Recently, multi-modal multi-label emotion recognition has aroused increasing interest. For example, (Ju et al. 2020; Zhang et al. 2021a) models modality-to-label and feature-to-label dependence besides label correlations.

Methodology

In this section, we first give the formulation of Multi-modal Multi-label Emotion Recognition (MMER). We use lowercase for scalars (e.g., a), uppercase for vectors (e.g., A) and bold for matrices (e.g., $\boldsymbol{a}$, $\boldsymbol{A}$). Let $\mathcal{X}^v = \mathbb{R}^{d_v \times \tau_v}$, $\mathcal{X}^a = \mathbb{R}^{d_a \times \tau_a}$, $\mathcal{X}^t = \mathbb{R}^{d_t \times \tau_t}$ be the visual (v), audio (a), text (t) feature space respectively, and $\mathcal{Y} = \{y_1, y_2, \cdots, y_l\}$ denote the label space with l labels, where $d_{\{v,a,t\}}$ represents modality dimension and $\tau_{\{v,a,t\}}$ represents sequence length. Given a training dataset $\mathcal{D} = \{(\boldsymbol{X}_i^{\{v,a,t\}}, Y_i)\}_{i=1}^n$ with n data samples, the goal of MMER is to learn a function $\mathcal{F} : \mathcal{X}^v \times \mathcal{X}^a \times \mathcal{X}^t \to 2^{\mathcal{Y}}$, which can assign a set of possible emotion labels for the unseen video. For the i-th video, $\boldsymbol{X}_i^{\{v,a,t\}} \in \mathcal{X}^{\{v,a,t\}}$ is the modality features and $Y_i \subseteq \mathcal{Y}$ is the set of relevant labels. Fig. 2 shows the main framework of TAILOR, which comprises the following modules: *Unimodal Extractor*, *Adversarial Multi-modal Refinement* and *Label-Modal Alignment*.

Uni-modal Extractor

The pre-extracted features for each modality in CMU-MOSEI (Zadeh et al. 2018c) dataset are represented by asynchronous coordinated sequences. To exploit long-term contextual information, we use n_v-layer, n_a-layer, n_t-layer Transformer encoder (Vaswani et al. 2017) to enrich the visual features, audio features and text features with sequence level context separately. The transformer encoder consists of two sub-layers: multi-head self-attention layer and position-wise feed-forward layer, where residual connections (He et al. 2016) are adopted, followed by layer normalization.
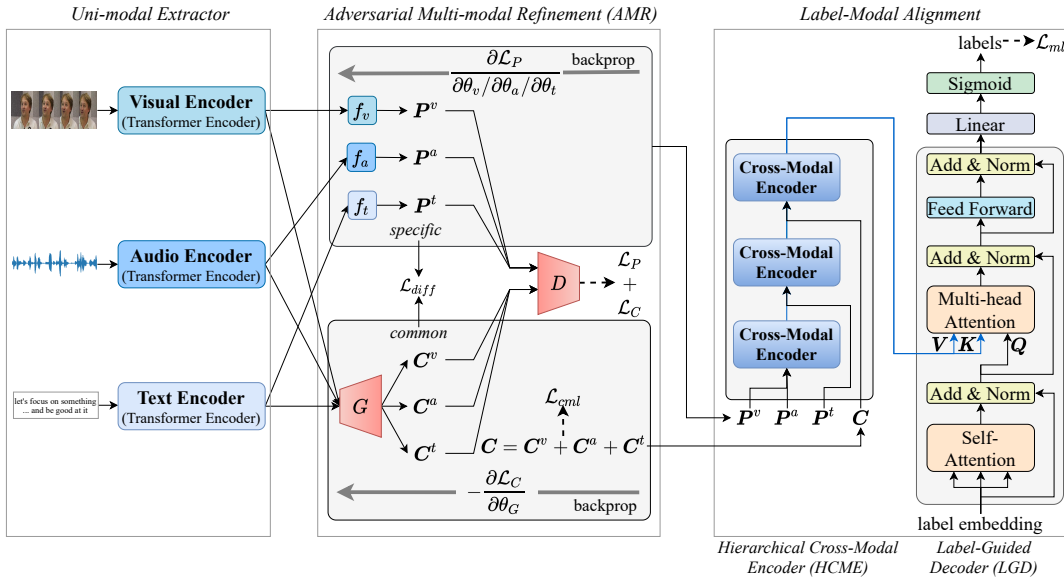
Figure 2: Overall structure of TAILOR.

As a result, we obtain new visual, audio and text embeddings, denoted as $\boldsymbol{V} \in \mathbb{R}^{d \times \tau}, \boldsymbol{A} \in \mathbb{R}^{d \times \tau}, \boldsymbol{T} \in \mathbb{R}^{d \times \tau}$, where d is the embedding dimension and τ is the sequence length.

Adversarial Multi-modal Refinement

It is well known that the greater the difference between inter-modal representations, the better the complementarity of inter-modal fusion (Yu et al. 2020). Though the uni-modal extractors capture long-term temporal context, they are unable to deal with feature redundancy due to modality gap. Inspired by adversarial networks (Goodfellow et al. 2014), we design an adversarial multi-modal refinement module for the subsequent fusion. It inherently decomposes multiple modalities to two disjoint parts: common and private representations so as to extract consistency and specificity of heterogeneous modalities collaboratively and individually.

To maintain consistency, we design a generator $G(\cdot; \theta_G)$ with parameters θ_G, to project different modalities into a common latent subspace with distributional alignment. Apart from commonality, each modality contains specific information, which can complement with other modalities. We adopt fully connected deep neural networks $f_v(\cdot; \theta_v)$, $f_a(\cdot; \theta_a)$ and $f_t(\cdot; \theta_t)$ with parameters $\{\theta_v, \theta_a, \theta_t\}$ to project uni-modal embedding $\{\boldsymbol{V}, \boldsymbol{A}, \boldsymbol{T}\}$ respectively. The common and private representations can be formulated as,

$$\begin{aligned} \boldsymbol{C}^v &= G(\boldsymbol{V}; \theta_G), \ \boldsymbol{C}^a = G(\boldsymbol{A}; \theta_G), \ \boldsymbol{C}^t = G(\boldsymbol{T}; \theta_G) \\ \boldsymbol{P}^v &= f_v(\boldsymbol{V}, \theta_v), \ \boldsymbol{P}^a = f_a(\boldsymbol{A}; \theta_a), \ \boldsymbol{P}^t = f_t(\boldsymbol{T}, \theta_t) \end{aligned} \quad (1)$$

where $\boldsymbol{C}^{\{v,a,t\}}, \boldsymbol{P}^{\{v,a,t\}} \in \mathbb{R}^{d \times \tau}$.

Adversarial Training In order to guarantee the purity of common and specific representations, we design a modality discriminator $D(, ; \theta_D)$ which maps the input $\boldsymbol{I} \in \mathbb{R}^{d \times \tau}$ into a probability distribution and estimates which modality the representation comes from, where d is the modality dimen-

sion and τ is the sequence length.

$$D(\boldsymbol{I}; \theta_D) = \text{softmax}(\boldsymbol{I}^T \boldsymbol{W} + \boldsymbol{b}) \quad (2)$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times 3}$ is the weight matrices, and $\boldsymbol{b} \in \mathbb{R}^{\tau \times 3}$ is the bias matrices. The ground truth modality label of $\boldsymbol{I}$ is denoted as $\boldsymbol{O} \in \{\boldsymbol{O}^v, \boldsymbol{O}^a, \boldsymbol{O}^t\}$,

$$\boldsymbol{O}^v = \begin{bmatrix} 1,0,0 \\ \cdots \\ 1,0,0 \end{bmatrix}, \boldsymbol{O}^a = \begin{bmatrix} 0,1,0 \\ \cdots \\ 0,1,0 \end{bmatrix}, \boldsymbol{O}^t = \begin{bmatrix} 0,0,1 \\ \cdots \\ 0,0,1 \end{bmatrix}, \quad (3)$$

where $\boldsymbol{O}^v, \boldsymbol{O}^a, \boldsymbol{O}^t \in \mathbb{R}^{\tau \times 3}$.

Common representations $\boldsymbol{C}^{\{v,a,t\}}$ are encoded in a shared latent subspace, which tends to be in the same distribution. Therefore, the generator $G(, ; \theta_G)$ are encouraged to confuse discriminator $D(, ; \theta_D)$ thus not to distinguish the source modality of $\boldsymbol{C}^{\{v,a,t\}}$. We reconstruct a training dataset $\mathcal{D}_C = \{(\boldsymbol{C}_i^v, \boldsymbol{O}^v), (\boldsymbol{C}_i^a, \boldsymbol{O}^a), (\boldsymbol{C}_i^t, \boldsymbol{O}^t)\}_{i=1}^n$ for common modality classification. The common adversarial loss is,

$$\mathcal{L}_C = -\frac{1}{n} \sum_{m \in \{v,a,t\}} \sum_{i=1}^n (\boldsymbol{O}^m \log(D(\boldsymbol{C}_i^m; \theta_D))) \quad (4)$$

where $\mathcal{L}_C$ is trained with gradient reversal layer (Ganin and Lempitsky 2015) that leaves the input unchanged during forward propagation and multiply the gradient by -1 during the backpropagation.

Private representations $\boldsymbol{P}^{\{v,a,t\}}$ are encoded in diverse latent subspaces, which tends to be in different distributions. Therefore, the discriminator $D(\cdot; \theta_D)$ are encouraged to distinguish the source of modality. We reconstruct a training dataset $\mathcal{D}_P = \{(\boldsymbol{P}_i^v, \boldsymbol{O}^v), (\boldsymbol{P}_i^a, \boldsymbol{O}^a), (\boldsymbol{P}_i^t, \boldsymbol{O}^t)\}_{i=1}^n$ for private modality classification. The private adversarial loss is,

$$\mathcal{L}_P = -\frac{1}{n} \sum_{m \in \{v,a,t\}} \sum_{i=1}^n (\boldsymbol{O}^m \log(D(\boldsymbol{P}_i^m; \theta_D))) \quad (5)$$

Orthogonal Constraint To encode different aspects of multi-modal data, we penalize redundancy in $\boldsymbol{C}^{\{v,a,t\}}$ and $\boldsymbol{P}^{\{v,a,t\}}$ with orthogonal loss as follows,

$$\mathcal{L}_{diff} = - \sum_{m\in\{v,a,t\}} \sum_{i=1}^{n} ||(\boldsymbol{C}_i^m)^T \boldsymbol{P}_i^m||_F^2 \quad (6)$$

where $|| \cdot ||_F^2$ is the squared Frobenius norm.

Common Semantics Although the common generator $G(,;\theta_G)$ and private extractors $f_v(,;\theta_v)$, $f_a(,;\theta_a)$ $f_t(,;\theta_t)$ are encouraged to encode different aspects of multi-modal information, they should exhibit the same semantics. We are motivated to design common semantic loss for multi-label classification with common representations $\boldsymbol{C}^{\{v,a,t\}}$,

$$\mathcal{L}_{cml} = - \sum_{m\in\{v,a,t\}} \sum_{i=1}^{n} \sum_{j=1}^{l} y_i^j \log\hat{y}_i^{j,m} + (1-y_i^j)\log(1-\hat{y}_i^{j,m})) \quad (7)$$

where $\hat{y}_i^{j,m}$ is predicted with $\boldsymbol{C}^m$ and y_i^j is the ground-truth. $y_i^j = 1$ if the j-th label is relevant, 0 otherwise.

Label-modal Alignment

After projecting into private and common representations respectively and collectively, we need to fuse them into a joint representation for multi-label classification.

Hierarchical Cross-Modal Encoder The refined common and private modality representations contain consistent and complementary information, while few or no information with regard to modality interactions. Simply concatenating them together ignores modality interactions, which might introduce redundant information and lead to suboptimal problem (Zhang et al. 2018). We propose a novel BERT-like (Kenton and Toutanova 2019) Cross-Modal Encoder to exploit modality interactions.
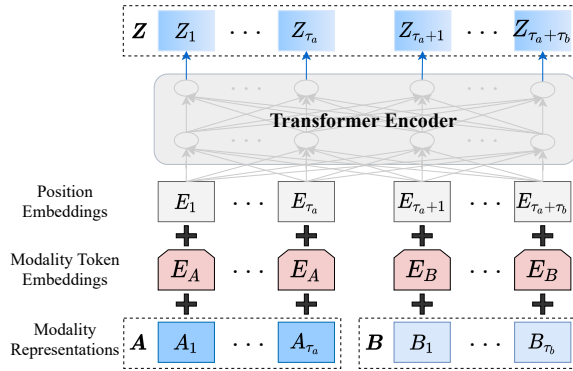

Figure 3: Structure of Cross-Modal Encoder (CME) between two modalities $\boldsymbol{A}$ and $\boldsymbol{B}$. The two input modality representations, modality token embeddings and position embeddings are summed and fed into the Transformer encoder.

Given two modalities a and b with representations $\boldsymbol{A} \in \mathbb{R}^{d\times\tau_a}$ and $\boldsymbol{B} \in \mathbb{R}^{d\times\tau_b}$, where d is modality dimension and $\tau_{\{a,b\}}$ is sequence length. On the one hand, to preserve the temporal information of the two modalities, we augment them with positional embeddings $\boldsymbol{E} \in \mathbb{R}^{d\times(\tau_a+\tau_b)}$. On the other hand, the feature distribution of various modalities are different due to heterogeneity, which poses a great challenge to multi-modal fusion. To bridge the large margin of the statistical properties between two modalities, we capture statistical regularities by adding two modality token embeddings $E_A \in \mathbb{R}^{1\times\tau_a}$ and $E_B \in \mathbb{R}^{1\times\tau_b}$ to modality a and b respectively. As illustrated in Fig. 3, the sum of modality representations, position embeddings and modality token embeddings is feed into n_c-layer Transformer Encoder, which outputs the joint representation $\boldsymbol{Z} \in \mathbb{R}^{d\times(\tau_a+\tau_b)}$ of modality a and b. Cross-Modal Encoder can be written as $\boldsymbol{Z} = \text{CME}(\boldsymbol{A}, \boldsymbol{B})$.

Besides, visual and audio modalities are more fine-grained than text modality in terms of granularity (Alayrac et al. 2020), which is rarely considered in existing fusion methods. To remedy the deficiency, we devise Hierarchical Cross-Modal Encoder (HCME) to exploit interactions across modalities with different level of granularities. Private representations $\boldsymbol{P}^{\{v,a,t\}}$ and common representations $\boldsymbol{C}^{\{v,a,t\}}$ are fused in a hierarchical structure and gradually complement with each other in a granularity descent way.

$$\begin{aligned} \text{Fine-grained} \quad & \boldsymbol{Z}^{va} = \text{CME}(\boldsymbol{P}^v, \boldsymbol{P}^a) \\ \text{Coarse-grained} \quad & \boldsymbol{Z}^{vat} = \text{CME}(\boldsymbol{Z}^{va}, \boldsymbol{P}^t) \quad (8) \\ \text{Mixed-grained} \quad & \boldsymbol{M} = \text{CME}(\boldsymbol{Z}^{vat}, \boldsymbol{C}) \end{aligned}$$

where $\boldsymbol{C} = \boldsymbol{C}^v + \boldsymbol{C}^a + \boldsymbol{C}^t \in \mathbb{R}^{d\times\tau}$, $\boldsymbol{M} \in \mathbb{R}^{d\times4\tau}$. HCME models 3 pairs of modalities with the fusion order $\psi = [v, a, t, c]$. Each pair of modalities interacts and correlates valuable information step by step.

Label-Guided Decoder Label correlations plays an important role in effective multi-label classification. For l possible labels in original label space $\boldsymbol{Y} = [Y_1, Y_2, \cdots, Y_n] \in \mathbb{R}^{l\times n}$, we use label indices to produce the label embedding $\boldsymbol{L} = [L_1; L_2; \cdots; L_l] \in \mathbb{R}^{l\times d}$ where l is the number of labels and d is label dimension which is the same as modality dimension. $\widetilde{k} = \{1, 2, \cdots, l\}\backslash k$ denotes all labels except the k-th label. $\boldsymbol{L}_k \in \mathbb{R}^{1\times d}$ is the label embedding for the k-th label, while $\boldsymbol{L}_{\widetilde{k}} = [L_1; \cdots; L_{k-1}; L_{k+1}; \cdots; L_l] \in \mathbb{R}^{(l-1)\times d}$ is the label embedding for $\widetilde{k}$. To exploit label correlations collaboratively, we adopt self-attention mechanism with h_l heads. For the i-th head [1],

$$\boldsymbol{q} = \boldsymbol{L}\boldsymbol{W}_i^q = \begin{bmatrix} \boldsymbol{L}_k \\ \boldsymbol{L}_{\widetilde{k}} \end{bmatrix} \boldsymbol{W}_i^q = \begin{bmatrix} \boldsymbol{q}_k \\ \boldsymbol{q}_{\widetilde{k}} \end{bmatrix}, \boldsymbol{W}_i^q \in \mathbb{R}^{d\times d/h_l} \quad (9)$$

Similarly, $\boldsymbol{k} = \begin{bmatrix} \boldsymbol{k}_k \\ \boldsymbol{k}_{\widetilde{k}} \end{bmatrix}, \boldsymbol{v} = \begin{bmatrix} \boldsymbol{v}_k \\ \boldsymbol{v}_{\widetilde{k}} \end{bmatrix}, \boldsymbol{W}_i^k, \boldsymbol{W}_i^v \in \mathbb{R}^{d\times d/h_l}$.

And then label correlation matrix $\boldsymbol{r}$ can be calculated as,

$$\boldsymbol{r} = \boldsymbol{q}\boldsymbol{k}^T = \begin{bmatrix} \boldsymbol{q}_k\boldsymbol{k}_k^T & \boldsymbol{q}_k\boldsymbol{k}_{\widetilde{k}}^T \\ \boldsymbol{q}_{\widetilde{k}}\boldsymbol{k}_k^T & \boldsymbol{q}_{\widetilde{k}}\boldsymbol{k}_{\widetilde{k}}^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_{kk} & \boldsymbol{r}_{k\widetilde{k}} \\ \boldsymbol{r}_{\widetilde{k}k} & \boldsymbol{r}_{\widetilde{k}\widetilde{k}} \end{bmatrix} \in \mathbb{R}^{l\times l} \quad (10)$$

where $\boldsymbol{r}_{kk}$ and $\boldsymbol{r}_{\widetilde{k}\widetilde{k}}$ represent the label-specific relation, $\boldsymbol{r}_{k\widetilde{k}}$ and $\boldsymbol{r}_{\widetilde{k}k}$ represent the interactive relation of the k-th label with respect to other $l-1$ labels. $\boldsymbol{r}_{k\widetilde{k}}$ denotes the influence

---

[1] we use bold lowercase to denote parameters in each head

of other $l-1$ labels to the k-th label, while $r_{\widetilde{k}k}$ denotes the influence of the k-th label to other $l-1$ labels. The label semantic embedding S of the i-th head is,

$$
\begin{aligned}
S_i = \begin{bmatrix} s_k \\ s_{\widetilde{k}} \end{bmatrix} &= \text{softmax}\left(\frac{1}{\sqrt{d/h_l}} \begin{bmatrix} r_{kk} & r_{k\widetilde{k}} \\ r_{\widetilde{k}k} & r_{\widetilde{k}\widetilde{k}} \end{bmatrix}\right) \begin{bmatrix} v_k \\ v_{\widetilde{k}} \end{bmatrix} \\
&= \begin{bmatrix} \sigma(r_{kk})v_k + \sigma(r_{k\widetilde{k}})v_{\widetilde{k}} \\ \sigma(r_{\widetilde{k}k})v_k + \sigma(r_{\widetilde{k}\widetilde{k}})v_{\widetilde{k}} \end{bmatrix} \in \mathbb{R}^{l \times d/h_l}
\end{aligned}
\tag{11}
$$

where $\sigma(r) = \text{softmax}(\frac{r}{\sqrt{d/h_l}})$ is a row-wise, scaled softmax. For the k-th label, the label-specific semantic embedding is $s_k = \sigma(r_{kk})v_k + \sigma(r_{k\widetilde{k}})v_{\widetilde{k}}$, which involves the collaboration of its own semantic implication and the semantic implication receiving from other labels. In addition, we add a residual connection followed by layer normalization (LN), to the final label-specific semantic embeddings,

$$
\begin{aligned}
L &= \text{LN}(L + S) \\
S &= \text{Concat}(S_1, S_2, \cdots, S_{h_l})W^L
\end{aligned}
\tag{12}
$$

where $W^L \in \mathbb{R}^{d \times d}$, $S \in \mathbb{R}^{l \times d}$.

Label semantics determine inherent dependence between labels and modalities. Therefore, the obtained label-specific semantic embeddings $L \in \mathbb{R}^{l \times d}$ can be further considered as a teacher to guide the learning of tailored representation for each label. Inspired by transformer decoder (Vaswani et al. 2017), we design a label-guided decoder to select discriminative information from joint multi-modal representations $M \in \mathbb{R}^{d \times 4\tau}$ with the guidance of label semantics. The latent dependence from modality space to label space is captured by multi-head attention with h_m heads,

$$
\begin{aligned}
Dep_{M \to L} &= \text{Concat}(\text{dep}_1, \cdots, \text{dep}_{h_m})W^M \\
\text{dep}_i &= \text{softmax}\left(\frac{LW_i^Q(M^TW_i^K)^T}{\sqrt{d/h_m}}\right)M^TW_i^V
\end{aligned}
\tag{13}
$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d/h_m}$, $\text{dep}_i \in \mathbb{R}^{l \times d/h_m}$, $W^M \in \mathbb{R}^{d \times d}$, $Dep_{M \to L} \in \mathbb{R}^{l \times d}$.

And then the tailored representations is generated by a standard feed-forward network (FFN) and two layer normalization (LN) with residual connection.

$$
H = \text{LN}(L + \text{FFN}(\text{LN}(L + Dep_{M \to L})))
\tag{14}
$$

where $H = [H_1; \cdots; H_l] \in \mathbb{R}^{l \times d}$.

Multi-label Classification For the k-th label, its tailored representation H_k is fed into a linear function followed by an output sigmoid for the final label classification,

$$
\mathcal{F}_k = \text{sigmoid}(H_kW_k + b_k)
\tag{15}
$$

where $W_k \in \mathbb{R}^d$ is weight vector and $b_k \in \mathbb{R}$ is the bias. The final multi-label classification loss can be computed with binary cross-entropy loss,

$$
\mathcal{L}_{ml} = -\sum_{i=1}^{n}\sum_{j=1}^{l} y_i^j \log \hat{y}_i^j + (1 - y_i^j)\log(1 - \hat{y}_i^j)
\tag{16}
$$

where $\hat{y}_i^j$ is predicted by Eq. 15 and y_i^j is the ground-truth. $y_i^j = 1$ if the j-th label is relevant, 0 otherwise.

Above all, combining the final multi-label classification loss $\mathcal{L}_{ml}$, common adversarial loss $\mathcal{L}_C$, private adversarial loss $\mathcal{L}_P$, common semantic loss $\mathcal{L}_{cml}$ and orthogonal loss $\mathcal{L}_{diff}$ together, the final objective function is computed as,

$$
\mathcal{L}_{All} = \mathcal{L}_{ml} + \alpha(\mathcal{L}_C + \mathcal{L}_P) + \beta\mathcal{L}_{diff} + \gamma\mathcal{L}_{cml}
\tag{17}
$$

where α, β and γ are the trade-off parameters.

Experiment

In this section, we give empirically evaluations and analysis of our proposed TAILOR method.

Experimental Setup

Dataset We conduct experiments on benchmark multi-modal multi-label dataset CMU-MOSEI [2] (Zadeh et al. 2018c), which contains $22,856$ video segments from $1,000$ distinct speakers. Each video inherently contains 3 modalities: visual, audio and text, while annotated with 6 discrete emotions: {*angry, disgust, fear, happy, sad, surprise*}. We pre-extract 35-dimensional visual features from video frames by FACET (Baltrušaitis, Robinson, and Morency 2016), 74-dimensional audio features from acoustic signals by COVAREP (Degottex et al. 2014) and 300-dimensional text features from video transcripts by Glove (Pennington, Socher, and Manning 2014). Table 1 summarizes details of CMU-MOSEI in both word-aligned and unaligned settings.

Table 1: Statistics of CMU-MOSEI, where $d_{\{v,a,t\}}$ is modality dimension and $\tau_{\{v,a,t\}}$ is sequence length.

	#instance	#train	#valid	#test	d_v	d_a	d_t	τ_v	τ_a	τ_t
aligned	22856	16326	1871	4659	35	74	300	60	60	60
unaligned								500	500	50

Evaluation Metrics We adopt 4 mostly used multi-label classification evaluation metrics (Zhang and Zhou 2014): Accuracy (Acc), Micro-F1, Precision (P) and Recall (R). Larger value indicates better performance.

Compared Approaches On the one hand, we conduct experiments with Multi-Label Classification (MLC) methods. For classic methods **BR** (Boutell et al. 2004), **LP** (Tsoumakas and Katakis 2007) and **CC** (Read et al. 2011), we concatenate all modalities as a new input. For text/image based methods, we only use text/image modality. **SGM** (Yang et al. 2018) views MLC as a sequence generation problem to take label correlations into account. **LSAN** (Xiao et al. 2019) considers document content and label texts simultaneously. **ML-GCN** (Chen et al. 2019b) captures label correlations for multi-label image recognition and employs GCN to map label representations.

On the other hand, we compare with multi-modal multi-label methods. **DFG** (Zadeh et al. 2018c) studies the nature of cross-modal dynamics in multimodal language. **RAVEN** (Wang et al. 2019) captures dynamic nature of nonverbal intents by shifting word representations based on the accompanying nonverbal behaviors. **MulT** (Tsai et al. 2019) fuses multi-modal information by directly attending

Table 2: Predictive performance of TAILOR on multi-modal multi-label CMU-MOSEI dataset with aligned and unaligned multi-modal sequences compared with state-of-the-arts. The best performance for each criterion is bolded.

Approaches	Aligned				Unaligned			
	Acc	P	R	Micro-F1	Acc	P	R	Micro-F1
BR (Boutell et al. 2004)	0.222	0.309	0.515	0.386	0.233	0.321	0.545	0.404
LP (Tsoumakas and Katakis 2007)	0.159	0.231	0.377	0.286	0.185	0.252	0.427	0.317
CC (Read et al. 2011)	0.225	0.306	0.523	0.386	0.235	0.320	0.550	0.404
SGM (Yang et al. 2018)	0.455	0.595	0.467	0.523	0.449	0.584	0.476	0.524
LSAN (Xiao et al. 2019)	0.393	0.550	0.459	0.501	0.403	0.582	0.460	0.514
ML-GCN (Chen et al. 2019b)	0.411	0.546	0.476	0.509	0.437	0.573	0.482	0.524
DFG (Zadeh et al. 2018c)	0.396	0.595	0.457	0.517	0.386	0.534	0.456	0.494
RAVEN (Wang et al. 2019)	0.416	0.588	0.461	0.517	0.403	0.633	0.429	0.511
MulT (Tsai et al. 2019)	0.445	0.619	0.465	0.531	0.423	0.636	0.445	0.523
SIMM (Wu et al. 2019)	0.432	0.561	0.495	0.525	0.418	0.482	0.486	0.484
MISA (Hazarika et al. 2020)	0.430	0.453	**0.582**	0.509	0.398	0.371	**0.571**	0.450
HHMPN (Zhang et al. 2021a)	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**

to low-level features in other modalities. **SIMM** (Wu et al. 2019) leverages shared subspace exploitation and view-specific information extraction with adversarial learning. **MISA** (Hazarika et al. 2020) learns modality-invariant and -specific representations as a pre-cursor to multi-modal fusion. **HHMPN** (Zhang et al. 2021a) simultaneously models feature-to-label, label-to-label and modality-to-label dependencies via graph message passing.

Implementation Details We set hyper-parameters $\alpha = 0.01$, $\beta = 5e{-}6$ and $\gamma = 0.5$. The batch size is 64. For layer number in Transformer Encoder, we set $n_v = n_a = 4$, $n_t = 6$ in uni-modal encoders, $n_c = 3$ in cross-modal encoders. The size of hidden layers in encoders and decoder is $d = 256$, the head number $h_l = h_m = 8$. All parameters in TAILOR are optimized by Adam (Kingma and Ba 2015) with an initial learning rate of $1e{-}5$ for aligned setting, $1e{-}4$ for unaligned setting and employ a liner decay learning rate schedule with a warm-up strategy. All experiments are running with one GTX 1080Ti GPU.

Experimental Results and Analysis

Experimental Results Except MulT, we include CTC (Graves et al. 2006) to be suitable for the unaligned setting. Based on the comparison results in Table 2, we have the following observations. 1) Our proposed TAILOR significantly surpasses the state-of-the art methods on all evaluation metrics except recall (R), which is relatively less important than accuracy (Acc) and Micro-F1 for performance evaluation. 2) **CC** performs best among 3 classic multi-label methods, which indicates the effectiveness of exploiting label correlations. 3) Text based multi-label methods **SGM**, **LSAN** and image based multi-label methods **ML-GCN** performs better than **CC**, which further conforms that label correlations conduce to capture more meaningful features. 4) **MulT** performs better than almost all multi-label methods that with concatenated modalities or only with text/image modality, which shows the necessity of exploiting modality complementarity. 5) Multi-modal multi-label methods such as **HHMPN** performs even better than aforementioned methods, which validates the effectiveness of exploiting modality information and label information simultaneously.

Ablation Study To get a better understanding of TAILOR, we investigate different components in 3 main modules: AMR, HCME, LGE. Ablation results are shown in Table 3

Table 3: Ablation experiments of TAILOR on the aligned CMU-MOSEI dataset. "w/o" denotes removing the component, "w/" denotes adding the component, ψ is fusion order. MTE: Modality Token Embedding, LE: Label Embedding, LC: Label Correlations. "w/ identical" denotes prediction with fused modalities via a dense layer.

Approaches	Acc	P	R	Micro-F1
(1) w/o AMR	0.446	0.634	0.474	0.543
(2) w/ $\mathcal{L}_{adv}$	0.432	0.722	0.419	0.530
(3) w/ $\mathcal{L}_{adv}$, $\mathcal{L}_{diff}$	0.462	0.581	0.520	0.549
(4) w/ $\boldsymbol{C}^{\{v,a,t\}}$	0.458	0.638	0.481	0.549
(5) w/ $\boldsymbol{P}^{\{v,a,t\}}$	0.449	0.605	0.496	0.545
(6) $\psi = [v, t, a, c]$	0.465	0.629	0.496	0.554
(7) $\psi = [a, t, v, c]$	0.470	0.584	0.524	0.552
(8) w/o MTE	0.478	0.601	0.528	0.562
(9) w/ identical	0.462	0.575	0.528	0.551
(10) w/ LE	0.465	0.558	**0.556**	0.557
(11) w/ LE, LC	0.473	0.594	0.538	0.564
(12) TAILOR	**0.488**	**0.641**	0.512	**0.569**

and several observations are obtained as follows.

- (1) is worst, which validates significance of adversarial multi-modal learning. As integrating $\mathcal{L}_{adv}$, $\mathcal{L}_{diff}$, $\mathcal{L}_{cml}$ to AMR optimization, (2), (3), (12) gradually improves.

- (4), (5) are worse than (12), which reveals that joint consideration of the commonality and diversity of multi-modal data leads to better performance.

- Changing the fusion order of HCME leads to poor performance, while (6), (7) is better than (1)-(5). It validates the rationality and optimality of HCME.

- (8) is worse than (12), where modality token embedding can really help bridge low-level modality gap.

- Label-specific results (10)-(12) gradually get better, which are all better than (9). Effective label-specific learning with label correlations and label-modal dependence can enhance discriminative power of each label.

Effect of Adversarial Learning In the Adversarial Mutimodal Refinement (AMR) module, we jointly optimize common adversarial loss, private adversarial loss, common semantic loss, orthogonal loss and overall loss. As shown in
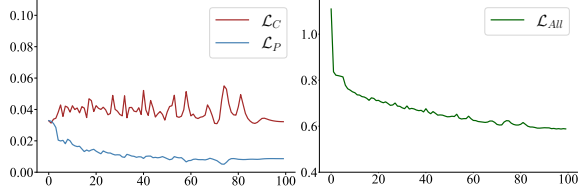

Figure 4: Common adversarial loss $\mathcal{L}_C$, private adversarial loss $\mathcal{L}_P$ and overall loss $\mathcal{L}_{All}$ w.r.t. the number of epochs.

Figure 4, private adversarial loss $\mathcal{L}_P$ and overall loss $\mathcal{L}_{All}$ decreases almost monotonously and converges smoothly, while common adversarial loss $\mathcal{L}_C$ first vibrates and later tends to be stable. In the end, it reaches a point where neither common nor private adversarial learning can improve.
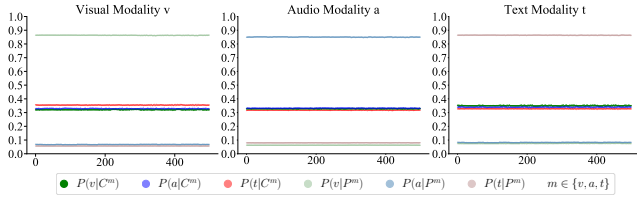

Figure 5: Distribution of private and common representations on different modalities. For modality $m \in \{v, a, t\}$,
$[P(v|\boldsymbol{C}^m), P(a|\boldsymbol{C}^m), P(t|\boldsymbol{C}^m)] = D(\boldsymbol{C}^m; \theta_D)$,
$[P(v|\boldsymbol{P}^m), P(a|\boldsymbol{P}^m), P(t|\boldsymbol{P}^m)] = D(\boldsymbol{P}^m; \theta_D)$.

Besides, we display the probability produced by discriminator $D(\cdot; \theta_D)$ in the AMR. In Fig. 5, for each modality m, the probabilities of common representations $P(v|\boldsymbol{C}^m), P(a|\boldsymbol{C}^m), P(t|\boldsymbol{C}^m)$ are centered around 0.33, which is hard to differentiate the source of common modalities. Contrarily, taking visual modality for example, $P(v|\boldsymbol{P}^m)$ is higher than $P(a|\boldsymbol{P}^m)$ and $P(t|\boldsymbol{P}^m)$ by a large margin, leading to increasingly separable representations.

Visualization of Learned Adversarial Representations t-SNE (Van der Maaten and Hinton 2008) is adopted to investigate the efficacy of adversarial multi-modal refinement. We visualize the common and private representations $\boldsymbol{C}^{\{v,a,t\}}$ and $\boldsymbol{P}^{\{v,a,t\}}$ learned without or with adversarial training and orthogonal constraint in aligned CMU-MOSEI. As shown in Fig. 6, the distributions of $\boldsymbol{C}^{\{v,a,t\}}$ and $\boldsymbol{P}^{\{v,a,t\}}$ are sometimes overlapped in the left subfigure. Contrarily, in the right subfigure, 1) the distributions of $\boldsymbol{C}^{\{v,a,t\}}$ are mixed together and increasingly blurred, where adversarial training proves effective to align distributions of different modalities and minimize the modality gap; 2) the common latent subspace is separable from each private subspace, where redundant latent representations are penalized with orthogonal constraint. In all, commonality and specificity of different modalities are well characterized.
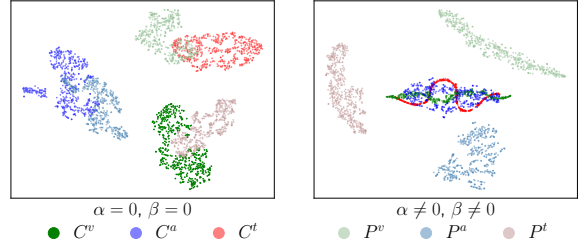

Figure 6: t-SNE visualization of common and private representations in the case without or with adversarial training. The green, blue, red colors represents visual, audio, text modalities respectively. And dark colors correspond to common parts, while light colors correspond to private parts.

Visualization of Learned Label Correlations We visualize the learned label correlations r in Eq. 11 to illustrate the interpretability. Due to page limit, we only exhibit the results in 4 attention heads. As shown in Figure 7, the learned
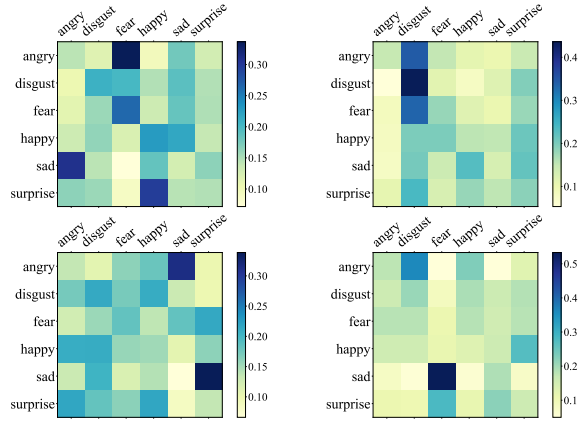

Figure 7: Label correlations visualization, indicating the influence of labels in each row to labels in each column. A higher blue intensity value indicates a stronger correlation.

label correlations differ from head to head, which jointly attends to rich semantic information from different perspectives. From the horizontal view, *angry* is highly correlated with *fear*, *disgust* and *sad* in different heads. In most cases, *surprise* is highly correlated with *happy*. All these correlations accord with our intuition.

Conclusion

In this paper, we propose versaTile multi-modAl learning for multI-labeL emOtion Recognition (TAILOR), consisting of uni-modal extractor, adversarial multi-modal refinement and label-modal alignment. These modules cooperate closely to refine private and common representations adversarially, fuse multiple modalities in terms of granularity gradually, and leverage label semantics to guide the construction of label-specific representation. Experimental results and analysis on both aligned and unaligned settings verify effectiveness and generalization of our proposed method.

Acknowledgments

References

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelovic, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-Supervised MultiModal Versatile Networks. *NeurIPS*, 2(6): 7.

Baltrusaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

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.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREPA collaborative voice analysis repository for speech technologies. In *ICASSP*, 960–964. IEEE.

Feng, L.; An, B.; and He, S. 2019. Collaboration based multi-label learning. In *AAAI*, volume 33, 3550–3557.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS*, 27.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 369–376.

Hazarika, D.; Zimmermann, R.; Poria; et al. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131.

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

Huang, J.; Li, G.; Huang, Q.; and Wu, X. 2016. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE transactions on knowledge and data engineering*, 28(12): 3309–3323.

Huang, Z.; Liu, F.; Wu, X.; Ge, S.; Wang, H.; Fan, W.; and Zou, Y. 2021. Audio-Oriented Multimodal Machine Comprehension via Dynamic Inter-and Intra-modality Attention. In *AAAI*, volume 35, 13098–13106.

Jiao, W.; Lyu, M.; and King, I. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *AAAI*, volume 34, 8002–8009.

Ju, X.; Zhang, D.; Li, J.; and Zhou, G. 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.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2020. Federated learning for vision-and-language grounding problems. In *AAAI*, volume 34, 11572–11579.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *ACL*, 2247–2256.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences. In *CVPR*, 2554–2562.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *AAAI*, volume 34, 164–172.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *AAAI*, volume 33, 6818–6825.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020a. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI*, volume 34, 1359–1367.

Mittal, T.; Guhan, P.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020b. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *CVPR*, 14234–14243.

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

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

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*, volume 33, 6892–6899.

Pham, H.; Manzini, T.; Liang, P. P.; and Poczós, B. 2018. Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis. *ACL 2018*, 53.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3): 333–359.

Saha, T.; Patra, A.; Saha, S.; and Bhattacharyya, P. 2020. Towards emotion-aided multi-modal dialogue act classification. In *ACL*, 4361–4372.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI*, volume 34, 8992–8999.

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.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning Factorized Multimodal Representations. In *ICLR*.

Tsoumakas, G.; and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3): 1–13.

Valada, A.; Mohan, R.; and Burgard, W. 2020. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5): 1239–1285.

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 *NeurIPS*, 5998–6008.

Wang, H.; Li, Z.; Huang, J.; Hui, P.; Liu, W.; Hu, T.; and Chen, G. 2020a. Collaboration Based Multi-Label Propagation for Fraud Detection. In *IJCAI*, 2477–2483.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020b. Deep multimodal fusion by channel exchanging. *NeurIPS*, 33.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, volume 33, 7216–7223.

Wu, X.; Chen, Q.-G.; Hu, Y.; Wang, D.; Chang, X.; Wang, X.; and Zhang, M.-L. 2019. Multi-View Multi-Label Learning with View-Specific Information Extraction. In *IJCAI*, 3884–3890.

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

Yang, P.; Sun, X.; Li, W.; Ma, S.; Wu, W.; and Wang, H. 2018. SGM: Sequence Generation Model for Multi-label Classification. In *COLING*, 3915–3926.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *ACL*, 3718–3727.

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.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. In *AAAI*, volume 32.

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

Zhang, C.; Yu, Z.; Hu, Q.; Zhu, P.; Liu, X.; and Wang, X. 2018. Latent semantic aware multi-view multi-label classification. In *AAAI*, 4414–4421.

Zhang, D.; Ju, X.; Zhang, W.; Li, J.; Li, S.; Zhu, Q.; and Zhou, G. 2021a. Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing. In *AAAI*, volume 35, 14338–14346.

Zhang, M.-L.; Fang, J.-P.; and Wang, Y.-B. 2021. BiLabel-Specific Features for Multi-Label Classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1): 1–23.

Zhang, M.-L.; and Wu, L. 2014. Lift: Multi-label learning with label-specific features. *IEEE transactions on pattern analysis and machine intelligence*, 37(1): 107–120.

Zhang, M.-L.; and Zhou, Z.-H. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge & Data Engineering*, 26(8): 1819.

Zhang, Q.-W.; Zhang, X.; Yan, Z.; Liu, R.; Cao, Y.; and Zhang, M.-L. 2021b. Correlation-Guided Representation for Multi-Label Text Classification. In *IJCAI*, 3363–3369.

Zhao, S.; Ma, Y.; Gu, Y.; Yang, J.; Xing, T.; Xu, P.; Hu, R.; Chai, H.; and Keutzer, K. 2020. An End-to-End visual-audio attention network for emotion recognition in user-generated videos. In *AAAI*, volume 34, 303–311.

Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2018. Multi-Label Learning with Global and Local Label Correlation. *IEEE Transactions on Knowledge & Data Engineering*, 30(06): 1081–1094.