The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

# Multi-modal Multi-label Emotion Recognition with Heterogeneous Hierarchical Message Passing

**Dong Zhang,**[1] **Xincheng Ju,**[1] **Wei Zhang,**[2] **Junhui Li,**[1]
**Shoushan Li,**[1] **Qiaoming Zhu,**[1] **Guodong Zhou**[1*]
[1] School of Computer Science and Technology, Soochow University, China
[2] Alibaba Group, China
dzhang@suda.edu.cn, xcju@stu.suda.edu.cn, zhangweinus@gmail.com, {jhli, lishoushan, qmzhu, gdzhou}@suda.edu.cn

## Abstract

As an important research issue in affective computing community, multi-modal emotion recognition has become a hot topic in the last few years. However, almost all existing studies perform multiple binary classification for each emotion with focus on complete time series data. In this paper, we focus on multi-modal emotion recognition in a multi-label scenario. In this scenario, we consider not only the label-to-label dependency, but also the feature-to-label and modality-to-label dependencies. Particularly, we propose a heterogeneous hierarchical message passing network to effectively model above dependencies. Furthermore, we propose a new multi-modal multi-label emotion dataset based on partial time-series content to show predominant generalization of our model. Detailed evaluation demonstrates the effectiveness of our approach.

## Introduction

Multi-modal Emotion Recognition has drawn more and more attention in natural language processing (Wang et al. 2019; Zhang et al. 2020a), speech analysis (Albanie et al. 2018; Priyasad et al. 2020), computer vision (Ruan et al. 2020; Deng et al. 2020) and multimedia analysis (Zhang et al. 2019a; Mai, Hu, and Xing 2020) communities. This is mainly due to not only its great facility to relevant tasks, such as emotional response generation and recommendation (Tsai et al. 2019), but also its wide applications, such as chatbot and sentiment systems (Chauhan et al. 2020).

Although a considerable amount of approaches have been proposed to handle multi-modal emotion recognition, there still exist two issues for this task:

**On the one hand**, conventional studies normally either perform multi-modal emotion recognition based on a dominant emotion or perform binary relevance for multiple emotions of a sample (Ju et al. 2020; Zhang et al. 2020a). However, an utterance of people or the contents of music tends to show multiple co-occurring emotions with potential correlations. Intuitively, multi-modal emotion recognition in a multi-label scenario brings greater challenges. In addition to various dependencies among emotion labels (label-to-label dependency), we should admit that the features from different modalities contribute differentially to each potential emotion (feature-to-label and modality-to-label dependencies). Therefore, we believe that a well-behaved approach should be unified to simultaneously model all dependencies at feature-to-label, modality-to-label and label-to-label levels for multi-modal multi-label emotion recognition (MMER).

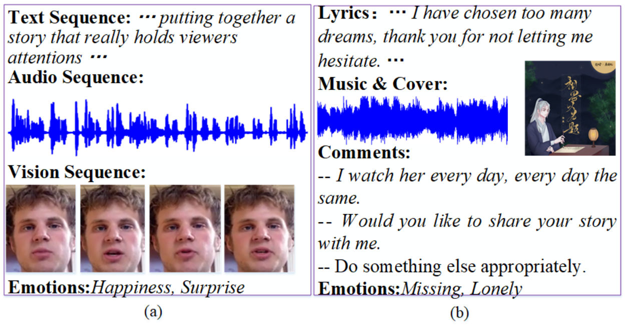
Text Sequence: ··· putting together a story that really holds viewers attentions ···
Audio Sequence:
Vision Sequence:
Emotions: Happiness, Surprise
(a)
Lyrics: ··· I have chosen too many dreams, thank you for not letting me hesitate. ···
Music & Cover:
Comments:
-- I watch her every day, every day the same.
-- Would you like to share your story with me.
-- Do something else appropriately.
Emotions: Missing, Lonely
(b)

Figure 1: Two examples of multi-modal multi-label (a) complete time series instance and (b) partial time series instance.

**On the other hand**, conventional studies normally employ sequential modeling (Zhang et al. 2020b) for multi-modal inputs (i.e., language, audio and video sequences) as shown in Figure 1(a). However, some utterances in multi-modal emotion analysis data may have no sequential relationships. As a result, the approaches with sequential modeling cannot be directly extended to this type of data, leading to performance loss. As shown in Figure 1(b), each piece of comment is independent, and there is no contextual relation. Therefore, we believe that a well-behaved approach should be applied to both complete time series data and partial time series data.

To handle above issues, we propose a novel heterogeneous hierarchical approach based on a neural message passing mechanism. The contributions of this paper can be summarized as follows:

- We propose a unified MMER approach, namely Hetero-

[*] Corresponding Author

geneous Hierarchical Message Passing Network (HHMPN), which can simultaneously model the feature-to-label, label-to-label and modality-to-label dependencies via graph message passing.

- We not only evaluate our approach on the existing multi-modal multi-label emotion dataset based on complete time series human utterances (MOSEI), but also propose a new multi-modal multi-label emotion dataset based on partial time series music (NEMu) for better evaluating the generalization of our model.

- Systematic experimentation on two benchmark datasets show that our approach can effectively address the challenges faced by MMER, and advance the state-of-the-art with a large margin.

## Related Work

As an interdisciplinary research field, emotion recognition becomes popular in both linguistic, non-linguistic and multi-modal disciplines (Zhang et al. 2019b). In the linguistic community, most existing studies of multi-label emotion recognition rely on special knowledge of emotion, such as context information (Li et al. 2015), cross domain (Yu et al. 2018) and external resource (Ying, Xiang, and Lu 2019). Besides, multi-label text classification approaches (Kim, Lee, and Jung 2018) are also employed for this task with the same principle. In the non-linguistic community, speech or visual features are normally used to perform multi-label emotion recognition with label dependence modeling (You et al. 2020). In the multi-modal community, related studies normally focus on single-label emotion task, whereas multi-label emotion task is typically limited to be transformed into multiple binary classification (Sun et al. 2020). In the following, we mainly overview multi-label emotion recognition and multi-modal emotion recognition.

**Multi-label Emotion Recognition.** 1) Linguistic: Recent studies normally cast multi-label emotion recognition task as a classification problem and leverage the special knowledge as auxiliary information (Ying, Xiang, and Lu 2019). These approaches may not be easily extended to those tasks without external knowledge. To this end, the multi-label text classification approaches can be quickly applied to emotion recognition (Chen et al. 2020; Fei et al. 2020). Recently, Yang et al. (2019) leverage a reinforced approach to find a better sequence than a baseline sequence, but it still relies on the pre-trained seq2seq model with a pre-defined order of ground-truth. 2) Non-linguistic: Ando et al. (2019) present a novel dominant emotion recognition method that improves on conventional hard-/soft-target based methods by directly handling the ambiguity of emotions for speech. Chen et al. (2019) propose a multi-label image classification model based on GCN. The model builds a directed graph over the object labels, where each node (label) is represented by word embeddings of a label, and GCN is learned to map this label graph into a set of inter-dependent object classifiers. This study is the most relevant to multi-label analysis with image information via graph-based network modeling.

Different from the above studies, we focus on multi-label emotion recognition in a multi-modal scenario by considering both feature-to-label and modality-to-label dependencies besides the label-to-label dependency.

**Multi-modal Emotion Recognition.** Recent studies on multi-modal emotion recognition largely depend on multi-modal fusion framework to perform binary classification within each emotion category. Chauhan et al. (2019) propose a contextual interactive attention network to learn the inter-modal interaction among the participating modalities. More recently, Mittal et al. (2020) uses canonical correlational analysis (CCA) to differentiate between an ineffectual and effectual input modality signal. However, they still handle the multi-label emotion recognition in a multi-modal scenario with binary relevance between one emotion and other emotions.

Different from the above studies, we focus on multi-modal emotion recognition in a multi-label scenario by considering label-to-label dependency besides feature-to-label and modality-to-label dependencies. Furthermore, we generalize our approach to newly collected partial time series data of MMER. To our best knowledge, it is the first attempt to explore this issue.

## Heterogeneous Hierarchical Message Passing Network for MMER

In this section, we introduce our approach for multi-modal multi-label emotion recognition, namely Heterogeneous Hierarchical Message Passing Network (HHMPN). In the following, we first give the background of message passing neural networks, then define the notations and introduce the specific functions of MP and UPD, finally give the details of four main hierarchical modules.

**Background.** Message Passing Neural Networks (MPNNs) (Gilmer et al. 2017) are a generalization of graph-based neural networks (GNN) (Scarselli et al. 2009). MPNNs model variables as nodes on a graph $\mathcal{G}$. Here $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ describes the set of nodes (variables) and $\mathcal{E}$ denotes the set of edges (about how variables interact with other variables). In an MPNN, joint representations of nodes and edges are modeled using message passing rather than explicit probabilistic formulations, allowing for efficient inference. MPNNs model the joint dependencies using message passing function $\mathrm{MP}^t$ and node updating function $\mathrm{UPD}^t$ for $T$ time steps, where $t$ is the current time step. The hidden state $v_i^t \in \mathbb{R}^d$ of node $i \in \mathcal{G}$ is updated based on messages $g_i^t$ from its neighboring nodes $\{v^t_{j\in\mathcal{N}(i)}\}$ defined by neighborhood $\mathcal{N}(i)$:

$$g_i^t = \sum_{j\in\mathcal{N}(i)} \mathrm{MP}^t(v_i^t, v_j^t) \quad (1)$$

$$v_i^{t+1} = \mathrm{UPD}^t(v_i^t, g_i^t) \quad (2)$$

After $T$ rounds of iterative updates to spread information to distant nodes, a readout function ROUT is used to form a specific vector on the updated node embeddings to make predictions like classifying nodes or classifying properties of the graph (Lanchantin, Sekhon, and Qi 2019).

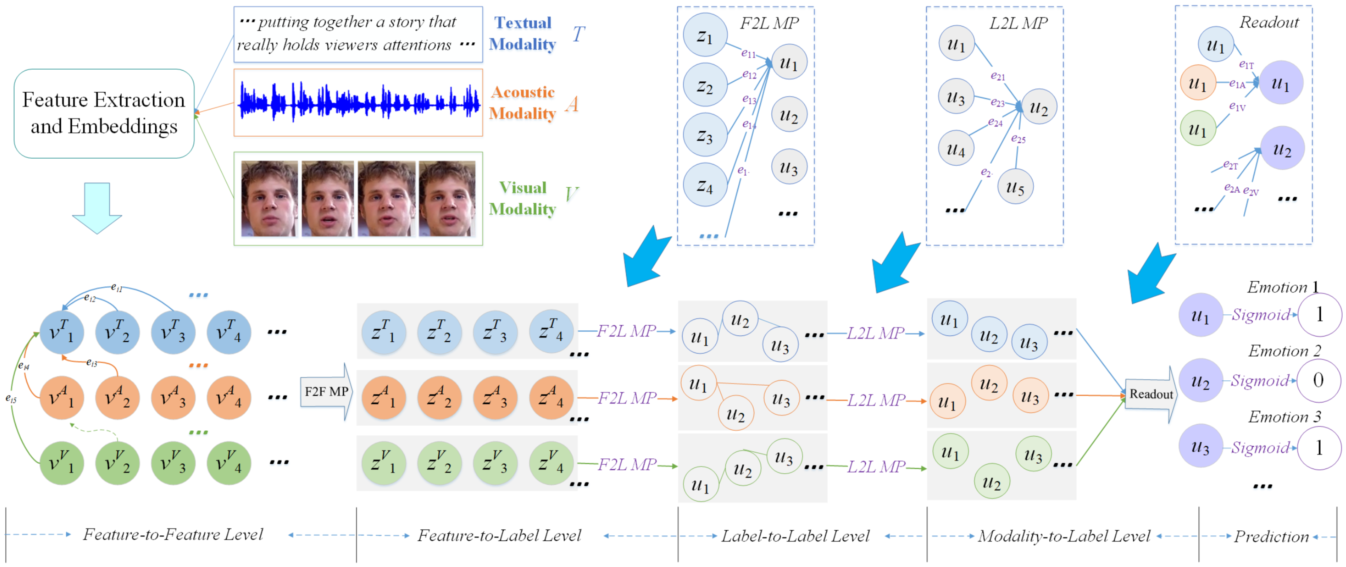

Figure 2: The overview of our proposed HHMPN ($T$=1). Note that 1) we take three modalities (Transcripts, Audio, and Facial expressions) in a video as an example, our model can support more modalities; 2) Our model can support both the sequential features and discrete features; 3) the feature number of different modalities in a sample tend to be different though we illustrate with the equal length.

## Definitions

We define the following notations, used throughout the paper. Let $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{N}$ be the set of data samples. Inputs $x^m \in \boldsymbol{x}$ are a set of $S^m$ components $\{x_1^m, x_2^m, \cdots, x_{S^m}^m\}$ of modality $m \in \mathcal{M}$, and outputs $\boldsymbol{y}$ are a set of $L$ labels $\{y_1, y_2, \cdots, y_L\}$, where $y_i \in \{0, 1\}$. In general, we can assume to represent the input feature components as embedded vectors $\{v_1^m, v_2^m, \cdots, v_{S^m}^m\}, v_i^m \in \mathbb{R}^d$ of modality $m$, using the learned embedding matrix $W_x^m \in \mathbb{R}^{\epsilon^m \times d}$. Here $d$ is the embedding size and $\epsilon^m$ is the size of $x_i^m$.

Similarly, labels can be first represented as embedded vectors $u_1^{t=0}, u_2^{t=0}, \cdots, u_L^{t=0}, u_i^t \in \mathbb{R}^d$, through a learned embedding matrix $W_y \in \mathbb{R}^{L \times d}$. Here we use $t$ to represent the 'state' of the embedding after the $t$-th update step. This is because in HHMPN network, each label embedding is updated for $t$ steps before the predictions are made. For different modalities, $u_i^{m,t}$ are all the same by initialization ($t = 0$) and next iteration.

**Message Passing Function.** We specifically choose self-attention for message passing, which enables nodes to attend over their neighborhoods differentially. This allows for the network to learn different importance for different nodes in a neighborhood, without depending on knowing the graph structure upfront. Formally,

$$\alpha_{ij}^t = softmax_j(e_{ij}^t) = \frac{exp(e_{ij}^t)}{\sum_{k \in \mathcal{N}(i)} exp(e_{ik}^t)} \quad (3)$$

$$e_{ij}^t = a(v_i^t, v_j^t) = \frac{(\boldsymbol{W}^{v1} v_i^t)^\top (\boldsymbol{W}^{v2} v_j^t)}{\sqrt{d}} \quad (4)$$

where $e_{ij}^t$ are normalized across all neighboring nodes of node $i$ using a softmax function. For the attention function $a(\cdot)$, we used a scaled dot product with node-wise linear transformations $W^{v1} \in \mathbb{R}^{d \times d}$ on node $v_i$ and $W^{v2} \in \mathbb{R}^{d \times d}$ on node $v_j$. Scaling by $\sqrt{d}$ is used to mitigate training issues.

Then, the message passing function in our approach can be defined as:

$$\mathrm{MP}_{att}(v_i^t, v_j^t; \boldsymbol{W}^1) = \alpha_{ij}^t \boldsymbol{W}^v v_j^t \quad (5)$$

$$g_i^t = v_i^t + \sum_{j \in \mathcal{N}(i)} \mathrm{MP}_{att}(v_i^t, v_j^t; \boldsymbol{W}^1) \quad (6)$$

where $g_i^t$ denotes the full message for node $v_i^t$ by linearly combining messages from all neighbor nodes $j \in \mathcal{N}(i)$ with a residual connection on current node. Message passing function is parameterized with matrices $\{\boldsymbol{W}^{v1}, \boldsymbol{W}^{v2}, \boldsymbol{W}^v\}$

**Updating Function.** Following (Lanchantin, Sekhon, and Qi 2019), node $v_i^t$ is updated to next state $v_i^{t+1}$ using message $g_i^t$ by a multi-layer perceptron (MLP) update function $\mathrm{UPD}_{mlp}$, plus a $g_i^t$ residual connection:

$$\mathrm{UPD}_{mlp}(g_i^t; \boldsymbol{W}^2) = \mathrm{ReLU}(\boldsymbol{W}^r g_i^t + b_1)^\top \boldsymbol{W}^b + b_2 \quad (7)$$

$$v_i^{(t+1)} = g_i^t + \mathrm{UPD}_{mlp}(g_i^t; \boldsymbol{W}^2) \quad (8)$$

where function $\mathrm{UPD}_{mlp}$ is parameterized with matrices $\{\boldsymbol{W}^r, \boldsymbol{W}^b\}$.

## Heterogeneous Hierarchical Message Passing Network

In this section, we mainly detail the core components of our HHMPN hierarchically.

**Feature-to-Feature Level.** In this module, we cast each extracted feature of each modality as a node in the graph

$\mathcal{G}_{xx} = \{\mathcal{V}_{xx}, \mathcal{E}_{xx}\}$. To model both the self- and cross-modal influence among multi-modal inputs, each node will be connected with other self-modal nodes and all cross-modal nodes, but separately as modality, which is shown in the "Feature-to-Feature Level" module of Figure 2. This feature-to-feature message passing (F2FMP) process can be defined as follows,

$$g_i^t = \bigoplus_{m\in\mathcal{M}} \sum_{j\in\mathcal{N}^m(i)} \mathrm{MP}_{att}(v_i^t, v_j^t; \boldsymbol{W}_{xx}^1) \tag{9}$$

$$z_i^t = g_i^t + \mathrm{UPD}_{mlp}(g_i^t; \boldsymbol{W}_{xx}^2) \tag{10}$$

where $\mathcal{M}$ denotes the set of multiple modalities (e.g., $\{Text, Vision, Audio\}$). $\mathcal{N}^m(i)$ denotes the neighbors of node $v_i^m$ in modality $m$. $\oplus$ denotes the concatenation operation. The weight $W^r$ in this process, i.e., $W_{xx}^r$ has the dimension of $d \times 3d$, the dimension of other weights are all the same as each module.

**Feature-to-Label Level.** For different modalities, we obtain the corresponding updated embedding $\{z_1^m, z_2^m, ..., z_{S^m}^m\}$ of modality $m$ from F2FMP with both self- and cross-modal interactions. On this basis, we cast each updated feature and each label as nodes in the graph $\mathcal{G}_{xy} = \{\mathcal{V}_{xy}, \mathcal{E}_{xy}\}$. Then, we set out to update the label embeddings by passing messages from the interactive multi-modal embeddings to the label embeddings, as shown in the "Feature-to-Label Level" module of Figure 2. Since both different features and different modalities contribute differently to a potential emotion label, we discriminatively perform feature-to-label message passing (F2LMP) as each modality. In this step, messages are only passed from the input nodes to the label nodes, and not vice versa (i.e., Feature-to-Label level message passing is directed). More specifically, to update label embedding $u_i^t$, we define as follows,

$$g_i^{m,t} = u_i^{m,t} + \sum_{j=1}^{S^m} \mathrm{MP}_{att}(u_i^{m,t}, z_j^m; \boldsymbol{W}_{xy}^1) \tag{11}$$

$$u_i^{m,t'} = g_i^{m,t} + \mathrm{UPD}_{mlp}(g_i^{m,t}; \boldsymbol{W}_{xy}^2) \tag{12}$$

**Label-to-Label Level.** In order to consider label dependencies, we model interactions between the label nodes $\{u_{1:L}^{t'}\}$ using label-to-label message passing (L2LMP) and update them accordingly, as shown in the "Label-to-Label Level" module of Figure 2. We assume there exist a label interaction graph $\mathcal{G}_{yy} = (\mathcal{V}_{yy}, \mathcal{E}_{yy})$, $\mathcal{V}_{yy} = y_{1:L}$, and $\mathcal{E}_{yy}$ includes all undirected pairwise edges connecting node $y_i$ and node $y_j$. This message passing process can be defined as follows,

$$g_i^{m,t} = u_i^{m,t'} + \sum_{j\in\mathcal{N}^m(i)} \mathrm{MP}_{att}(u_i^{m,t'}, u_j^{m,t'}; \boldsymbol{W}_{yy}^1) \tag{13}$$

$$u_i^{m,t+1} = g_i^{m,t} + \mathrm{UPD}_{mlp}(g_i^{m,t}; \boldsymbol{W}_{yy}^2) \tag{14}$$

where $\mathcal{N}^m(i)$ denotes the label neighbors of label node $i$ for modality $m$.

**Modality-to-Label Level.** To handle the modality-to-label dependency, we adaptively combine label embeddings from different modalities. Specifically, we leverage modality-to-label readout attention to learn the different importance of each modality, as shown in the "Modality-to-Label Level" module of Figure 2. Formally,

$$u_i^{t+1} = \sum_{m\in\mathcal{M}} \alpha_i^m u_i^{m,t+1} \tag{15}$$

$$\alpha_i^m = \frac{exp(\boldsymbol{W}^l u_i^{m,t+1})}{\sum_{m'\in\mathcal{M}} exp(\boldsymbol{W}^l u_i^{m',t+1})} \tag{16}$$

## Loss Function

After $T$ updates to the label embeddings, the last module predicts each label $\{\hat{y}_1, \cdots, \hat{y}_L\} = \hat{\boldsymbol{y}}$. A prediction function projects each of the $L$ label embeddings $u_i^T$ using projection matrix $W^o \in \mathbb{R}^{L\times d}$, where row $W_i^o \in \mathbb{R}^d$ is the learned output vector for label $i$. The calculated vector of size $L \times 1$ is then fed through an element-wise sigmoid function to produce probabilities of each label being positive:

$$\hat{y}_i = \mathrm{sigmoid}(W_i^o u_i^T) \tag{17}$$

The final output of our model is trained using the mean binary cross entropy (BCE) over all targets. For one sample, given true binary label vector $\boldsymbol{y}$ and predicted labels $\hat{\boldsymbol{y}}$, the output loss is:

$$\mathcal{J}_{out} = \frac{1}{L}\sum_{i=1}^{L} -(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \tag{18}$$

We also leverage an auxiliary loss item to explore the label nodes at each intermediate state from $t = 1$ to $T-1$. Specifically, we use the same matrix $W^o$ to extract the intermediate prediction $\hat{y}^t$ at state $t$. Then, we use the same BCE loss on these predictions to compute intermediate loss:

$$\mathcal{J}_{int} = \frac{1}{L}\sum_{i=1}^{L} -(y_i \log(\hat{y}_i^t) + (1-y_i)\log(1-\hat{y}_i^t)) \tag{19}$$

Therefore, the final loss is a combination of both the original and intermediate, where the intermediate loss is weighted by $\lambda$:

$$\mathcal{J}_{all} = \mathcal{J}_{out} + \lambda \sum_{t=1}^{T-1} \mathcal{J}_{int} \tag{20}$$

# Experimentation

## Experimental Settings

**Dataset.** 1) MOSEI is the only public benchmark for MMER in English. The document-level videos of this dataset are segmented into utterances with three modalities, i.e., the textual, visual and acoustic modalities, while the emotion categories contain *happiness*, *sadness*, *anger*, *fear*, *disgust* and *surprise*. 2) To further demonstrate the generalization of our approach, we collect a partial time series dataset for MMER from NetEase Cloud Music[1], namely NEMu. Each sample of this dataset includes lyric, comments, audio, images (e.g., cover and posters) and meta-data,

[1] music.163.com

| Dataset | Split | | | Multi-label | | Other | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Train | Valid | Test | One | Two and more | Avg. Time(s) | Language | Labels | Avg. Words/Sentence |
| MOSEI | 16326 | 1871 | 4659 | 11121 | 8339 | 7.28/Utterance | English | 6 | 19.1 |
| NEMu | 15125 | 1891 | 1891 | 5587 | 13320 | 115/Sample | Chinese | 12 | 24.0 |

Table 1: The statistics summary on the CMU-MOSEI and NEMu datasets.

as well as the corresponding emotion label set[2], such as *Sad*, *Excited* and *Happy*, etc. In our settings, we mainly use lyrics, comments, audio and images, which are considered as four different modalities. The statistics of these two datasets are shown in Table 1.

**Feature Extraction and Embeddings.** For MOSEI, we refer to the original paper[3] (Zadeh et al. 2018). For NEMu, lyrics and comments modalities are pre-processed with glove embeddings ($\epsilon^L$ and $\epsilon^C$ = 300). The audio modality is pre-processed with Librosa for MFCCs ($\epsilon^A$ = 74). The image modality is pre-processed with ResNet (He et al. 2016) ($\epsilon^V$ = 2048).

**Implementation Details.** For both datasets, we use the same hyper-parameters: the size $d$ of the hidden layer in each modality is 256, iteration times $T$ is set 3, batch size is 64 and $\lambda$ in joint loss is 0.2. We train HHMPN in an end-to-end manner by minimizing the joint loss function with the Adam optimizer (Kingma and Ba 2015). Besides, we make use of the dropout regularization (Srivastava et al. 2014) to avoid overfitting and clip the gradients (Pascanu, Mikolov, and Bengio 2013) to the maximum norm of 10.0. During training, we train each model for a fixed number of epochs 50 and monitor its performance on the validation set. Once the training is finished, we select the model with the best $F_1$ score on the validation set as our final model and evaluate its performance on the test set. For a better comparison, we perform 10 fold cross-validation in all our experiments. To motivate future research, both code and dataset will be released[4].

**Evaluation Metrics.** In our study, we employ four evaluation metrics to measure the performance of different approaches to MMER, i.e., multi-label Accuracy (*Acc*), Hamming Loss (*HL*), micro $F_1$ measure ($miF_1$), macro $F_1$ measure ($maF_1$). These metrics have been popularly used in some multi-label classification problems (Li et al. 2015; Wu, Xiong, and Wang 2019; Lanchantin, Sekhon, and Qi 2019; Ma et al. 2020). Note that smaller *HL* corresponds to better classification quality, while larger *Acc*, $miF_1$ and $maF_1$ correspond to better classification quality.

## Baselines

To our best knowledge, there are no unified multi-modal multi-label approaches for emotion recognition. Therefore, we mainly compare the uni-modal multi-label emotion recognition approaches, which early fuse the multi-modal sequences as new input, and multi-modal single-label emotion recognition approaches, which leverage multi-modal fusion to generate a final representation and use a linear layer of $L$ dimensions with $sigmoid$ activation to predict the emotions.

The first group can be also splitted as the classical, linguistic and non-linguistic multi-label approaches. Classical: (1) **BR** (Shen et al. 2004), binary relevance, which ignores the correlations between labels. (2) **CC** (Read et al. 2011), classifier chain, which considers high-order label correlations. (3) **LP** (Tsoumakas, Katakis, and Vlahavas 2011), which breaks the initial set of labels into a number of small random subsets and training a corresponding classifier. Linguistic multi-label: (4) **LSAN** (Xiao et al. 2019), which takes advantage of label semantic information to determine the semantic connection between labels and documents for constructing label-specific document representation. (5) **Seq2Set** (Yang et al. 2019), which leverages deep reinforcement learning to find a most probable sequence as the target label set based on a pre-trained sequence-to-sequence model of RNN. (6) **KRF** (Ma et al. 2020), which jointly exploits the inherent relations between music styles according to external knowledge and their statistical relations by review-driven modeling. Non-linguistic multi-label: (7) **ML-GCN** (Chen et al. 2019), which predicts a set of object labels that present in an image in a task of multi-label image recognition. As objects normally co-occur in an image, it is desirable to model the label dependencies to improve the recognition performance. (8) **MLEE** (Ando et al. 2019), which adopts a multi-task framework to jointly predict multi-label emotions and dominant emotion, which provides the auxiliary information for each other. Thus, this approach relies on the annotated dominant emotion, which can not be obtained in NEMu dataset. This reason explains why no results on NEMu for this approach.

The second group mainly focus on multi-modal fusion. (9) **MulT** (Tsai et al. 2019), which addresses the issues about inherent data non-alignment due to variable sampling rates for the sequences from each modality and long-range dependencies between elements across modalities in an end-to-end manner without explicitly aligning the data. This approach is considered as the state-of-the-art in multi-modal emotion recognition. (10) **CIA** (Chauhan et al. 2019), which learns the inter-modal interaction among the participating modalities through an auto-encoder mechanism and employs a context-aware attention module to exploit the correspondence among the neighboring utterances. (11) **M3ER** (Mittal et al. 2020), which models a data-driven multiplicative fusion method to emphasize the more reliable cues and suppress others on a per-sample basis. This approach is con-

[2] Although there are other labels of music style, we just retain the samples only containing emotion labels.

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

[4] https://github.com/MANLP-suda/HHMPN

| Approach | | MOSEI | | | | NEMu | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Acc*(%) | *HL* | $miF_1(\%)$ | $maF_1(\%)$ | *Acc*(%) | *HL* | $miF_1(\%)$ | $maF_1(\%)$ |
| Classical | BR (Shen et al. 2004) | 22.2 | 0.371 | 38.6 | 34.7 | 23.0 | 0.457 | 41.1 | 40.5 |
| | CC (Read et al. 2011) | 22.5 | 0.377 | 38.6 | 34.1 | 23.5 | 0.465 | 41.7 | 41.1 |
| | LP (Tsoumakas et al. 2011) | 15.9 | 0.426 | 28.6 | 28.8 | 21.1 | 0.414 | 37.2 | 35.0 |
| Linguistic | LASN (Xiao et al. 2019) | 39.3 | 0.209 | 50.1 | 32.3 | 19.5 | 0.332 | 39.7 | 35.7 |
| | Seq2set (Yang et al. 2019) | 45.7 | 0.231 | 53.8 | 34.0 | 24.8 | 0.424 | 42.1 | 39.7 |
| | KRF (Ma et al. 2020) | 45.3 | 0.226 | 51.5 | 29.0 | 23.1 | 0.496 | 42.0 | 39.7 |
| Non-linguistic | ML-GCN (Chen et al. 2019) | 41.1 | 0.207 | 50.9 | 29.7 | 15.8 | 0.293 | 34.4 | 27.8 |
| | MLEE (Ando et al. 2019) | 43.7 | 0.211 | 52.8 | 38.6 | - | - | - | - |
| Multi-modal | MulT (Tsai et al. 2019) | 44.5 | 0.190 | 53.1 | 34.4 | 17.9 | 0.293 | 42.6 | 39.0 |
| | CIA (Chauhan et al. 2019) | 42.9 | 0.214 | 45.5 | 11.7 | 11.1 | 0.336 | 29.6 | 34.0 |
| | M3ER (Mittal et al. 2020) | 40.9 | 0.195 | 51.9 | 34.9 | 19.4 | 0.281 | 40.6 | 36.4 |
| | HHMPN (Ours) | **45.9** | **0.189** | **55.6**$^{\dagger\ddagger}$ | **43.0** $^{\dagger\ddagger}$ | **24.9** | **0.270** | **46.1**$^{\dagger\ddagger}$ | **43.5**$^{\dagger\ddagger}$ |
| | HHMPN w/o F2F | 41.1 | 0.190 | 52.2 | 33.6 | 22.3 | 0.341 | 43.1 | 39.5 |
| | HHMPN w/o F2L | 44.8 | 0.204 | 53.5 | 35.4 | 23.7 | 0.370 | 43.8 | 41.8 |
| | HHMPN w/o L2L | 41.9 | 0.323 | 53.5 | 35.7 | 19.6 | 0.326 | 42.4 | 39.5 |
| | HHMPN w/o M2L | 44.0 | 0.232 | 54.3 | 40.2 | 24.6 | 0.359 | 45.4 | 43.1 |

Table 2: Performance of different approaches to MMER on both MOSEI and NEMu datasets. Note that 1) although the classical, linguistic and non-linguistic multi-label approaches are originally designed for uni-modal data, we early fuse the multi-modal data as a new input in our implementation for a fair comparison. 2) KRF and ML-GCN support the non-time series data as input, which are the representative baselines for non-sequential modeling. The marker † refers to significant test $p\text{-}value < 0.05$ when comparing with MulT, and the marker ‡ refers to significant test $p\text{-}value < 0.05$ when comparing with M3ER.

sidered as the state-of-the-art on MOSEI dataset.

## Experimental Results

**Comparison with Representative Baselines.** Table 2 shows the performance of different approaches to multi-modal multi-label emotion recognition on both MOSEI and NEMu datasets. From this table, we can see that: First, recent linguistic and non-linguistic multi-label approaches outperform classical multi-label approaches in most cases on MOSEI, they do not show obvious advantages on NEMu and are even inferior to the classical multi-label approaches in some metrics. For example, regarding $miF_1$, **LASN** performs better than **CC** by 1.15% on MOSEI, but obviously worse than **CC** on NEMu. This suggests that there is a difference between temporal and partial temporal data. Second, compared with uni-modal multi-label approaches, multi-modal approaches typically perform better, especially **MulT**. This indicates multi-modal data need to well model interactions among different modalities. Third, although partial temporal data can be roughly fused into a complete sequential (**LASN**) or non-sequential (**ML-GCN**) model, the performance shows a worse trend. This is mainly because the existing models are not general to these two types of data. Finally, among all approaches, our **HHMPN** performs best and significantly better than SOTA in both multi-label and multi-modal emotion recognition areas.

**Ablation Study.** We also implement the ablated approaches without one specific component of our full model. 1) **HHMPN w/o F2F**, a variation of our approach, which removes feature-to-feature level message passing and retains the other modules. 2) **HHMPN w/o F2L**, a variation of our approach, which removes feature-to-label level message passing and retains the other modules. 3) **HHMPN w/o L2L**, a variation of our approach, which removes label-to-label level message passing and retains the other modules. 4) **HHMPN w/o M2L**, a variation of our approach, which replaces the modality-to-label readout with sum readout and retains the other modules.

As shown in Table 2, no matter which module is removed, the performance of model variants shows varying degrees of degradation. This suggests the importance of our approach consisting of these four modules. In addition, the importance of the four modules is different. For example, the performance degradation caused by **HHMPN w/o F2F** is the most, which indicates that effective multi-modal interactions are the most important in MMER. With respect to **HHMPN w/o L2L**, it shows the second rank of a performance drop, which verifies the correctness of our motivation.

## Analysis and Discussion

Since single-modal, bi-modal, visualization and other analysis on MOSEI has been well investigated by previous studies (Chauhan et al. 2019; Mittal et al. 2020), we mainly provides further analysis of partial time series dataset NEMu.

**Case Study and Visualization.** Figure 3 illustrates a case that visualizes the importance of features (sentences) in comments passing to label nodes. From this figure, we can see that the fifth sentence "5) *although I am single and weak* $\cdots$" plays a most important role for label *Sad*. Besides, the second sentence from the bottom "14) $\cdots$ no longer related to me $\cdots$" conveys the most information about emotion *Lonely*. These learned attention are consistent with human decision, which suggests that our F2LMP module can effectively pass important information into different label representations.

Figure 4 visualizes the dependency degree among different labels. We can observe that *Miss* and *Sad*, *Relaxed* and *Happy* co-exist with large probabilities learned by our

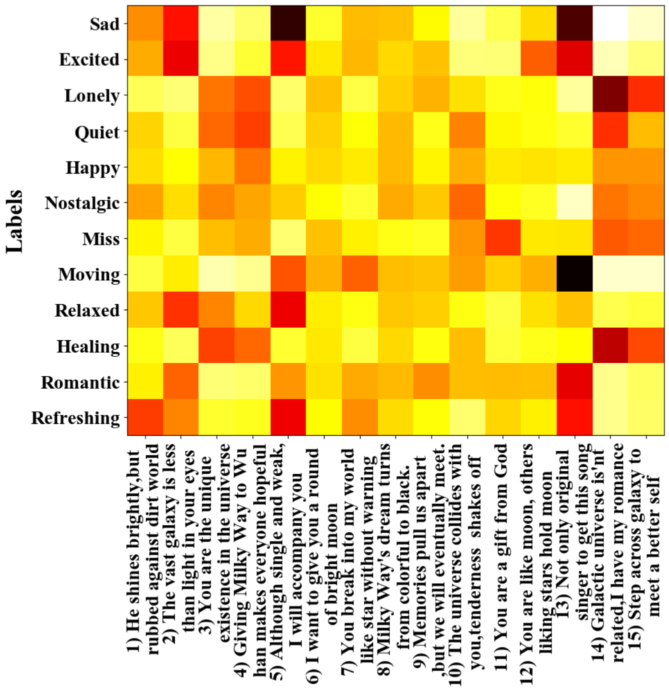

Figure 3: Visualization of F2LMP.

| Modalities | $Acc(\%)$ | $HL$ | $miF_1(\%)$ | $maF_1(\%)$ |
|---|---|---|---|---|
| L-A-C | 20.9 | 0.297 | 43.2 | 39.2 |
| L-A-V | 20.9 | 0.342 | 38.9 | 35.6 |
| L-C-V | 20.5 | 0.284 | 41.7 | 33.1 |
| A-C-V | 18.7 | 0.306 | 39.4 | 34.7 |
| L-A | 17.4 | 0.310 | 36.0 | 33.0 |
| L-C | 20.1 | 0.305 | 38.8 | 35.9 |
| L-V | 19.9 | 0.348 | 37.8 | 32.5 |
| A-C | 18.1 | 0.310 | 38.2 | 35.2 |
| A-V | 20.8 | 0.353 | 37.4 | 34.1 |
| C-V | 20.0 | 0.333 | 37.9 | 35.0 |
| L | 15.0 | 0.311 | 33.9 | 28.8 |
| A | 13.8 | 0.276 | 30.7 | 22.2 |
| C | 15.1 | 0.297 | 35.1 | 30.5 |
| V | 19.9 | 0.346 | 34.8 | 30.7 |

Table 3: Performance of uni-modality, bi-modality or tri-modality as input into our approach on NEMu dataset. L: lyrics, C: comments, V: vision, A: audio.

model, which indicates the effectiveness of label-to-label message passing.

**Impact of Less Modalities.** To further valid the necessity of our multi-modal approach in a multi-label scenario, we also report whether the lack of one or more modalities has a great impact on emotion recognition, as shown in Table 3. From this table, we can observe that no matter which one or more modalities are removed, the performance of multi-label emotion recognition will decrease to varying degrees. This suggests that it is necessary to propose a proper modality-aware approach for multi-label emotion recognition.

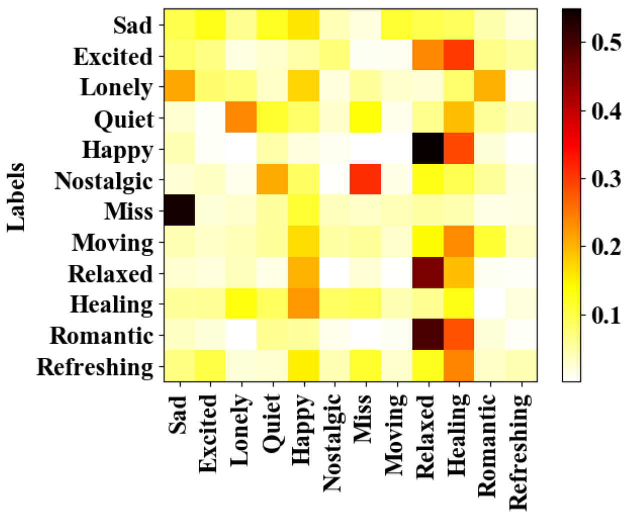

Figure 4: Visualization of L2LMP.

| **Most Emotions** | **% of Samples** | **Micro F1**(%) |
|---|---|---|
| Healing | 43.9 | 58.9 |
| Miss | 35.0 | 54.9 |
| Moving | 31.0 | 55.7 |
| **Fewest Emotions** | **% of Samples** | **Micro F1**(%) |
| Happy | 22.8 | 33.8 |
| Relaxed | 22.4 | 36.8 |
| Excited | 14.4 | 20.9 |

Table 4: The performance of the proposed approach on the most and fewest emotion labels for NEMu dataset.

**Error Analysis.** Although the proposed **HHMPN** has achieved significant improvements, we also notice that there are some limitations. For example, the proposed approach performs worse on the emotions with low frequency in the training set. Table 4 compares the performance on the top 3 music emotions of the highest and lowest frequencies. As we can see, the top 3 fewest emotions get much worse results than the top 3 most emotions. This is because the label distribution is highly imbalanced where unpopular emotions in music have too little training data.

## Conclusion

To handle the challenges faced by multi-modal multi-label emotion recognition, this paper proposes a heterogeneous hierarchical message passing network (HHMPN). This approach can be easily performed on both complete time series data and partial time series data. Besides, the dependency issues of feature-to-feature, feature-to-label, label-to-label and modality-to-label can be well captured through the core modules in our approach. Furthermore, we collect a new multi-modal multi-label dataset of partial time series for better evaluating our approach.

## Acknowledgments

We thank our anonymous reviewers for their valuable suggestions. This work was supported by a Major Project of Ministry of Science and Technology of China No.2020AAA0108604 and an NSFC grant No.62076176. This work was also supported by a project funded by China Postdoctoral Science Foundation No.2020M681713, a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and a joint research project of Alibaba Group and Soochow University.

## References

Albanie, S.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In *Proceedings of ACM MM 2018*, 292–301.

Ando, A.; Masumura, R.; Kamiyama, H.; Kobashikawa, S.; and Aono, Y. 2019. Speech Emotion Recognition Based on Multi-Label Emotion Existence Model. In *Proceedings of INTERSPEECH 2019*, 2818–2822.

Chauhan, D. S.; Akhtar, M. S.; Ekbal, A.; and Bhattacharyya, P. 2019. Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis. In *Proceeding of EMNLP-IJCNLP 2019*, 5646–5656.

Chauhan, D. S.; R, D. S.; Ekbal, A.; and Bhattacharyya, P. 2020. Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In *Proceedings of ACL 2020*, 4351–4360.

Chen, B.; Huang, X.; Xiao, L.; and Jing, L. 2020. Hyperbolic Capsule Networks for Multi-Label Classification. In *Proceedings of ACL 2020*, 3115–3124.

Chen, Z.; Wei, X.; Wang, P.; and Guo, Y. 2019. Multi-Label Image Recognition With Graph Convolutional Networks. In *Proceedings of IEEE-CVPR 2019*, 5177–5186.

Deng, D.; Chen, Z.; Zhou, Y.; and Shi, B. E. 2020. MIMAMO Net: Integrating Micro- and Macro-Motion for Video Emotion Recognition. In *Proceedings of AAAI 2020*, 2621–2628.

Fei, H.; Zhang, Y.; Ren, Y.; and Ji, D. 2020. Latent Emotion Memory for Multi-Label Emotion Classification. In *Proceedings of AAAI 2020*, 7692–7699.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural Message Passing for Quantum Chemistry. In *Proceedings of ICML 2017*, 1263–1272.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR 2016*, 770–778.

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 ACM MM 2020*, 512–520.

Kim, Y.; Lee, H.; and Jung, K. 2018. AttnConvnet at SemEval-2018 Task 1: Attention-based Convolutional Neural Networks for Multi-label Emotion Classification. In *Proceedings of SemEval@NAACL-HLT 2018*, 141–145.

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

Lanchantin, J.; Sekhon, A.; and Qi, Y. 2019. Neural Message Passing for Multi-label Classification. In *Proceedings of ECML-PKDD 2019*, 138–163.

Li, S.; Huang, L.; Wang, R.; and Zhou, G. 2015. Sentence-level Emotion Classification with Label and Context Dependence. In *Proceedings of ACL 2015*, 1045–1053.

Ma, Q.; Yuan, C.; Zhou, W.; Han, J.; and Hu, S. 2020. Beyond Statistical Relations: Integrating Knowledge Relations into Style Correlations for Multi-Label Music Style Classification. In *Proceedings of WSDM 2020*, 411–419.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *Proceedings of AAAI 2020*, 164–172.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020. M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues. In *Proceedings of AAAI 2020*, 1359–1367.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of ICML 2013*, 1310–1318.

Priyasad, D.; Fernando, T.; Denman, S.; Sridharan, S.; and Fookes, C. 2020. Attention Driven Fusion for Multi-Modal Emotion Recognition. In *Proceedings of ICASSP 2020*, 3227–3231.

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

Ruan, S.; Zhang, K.; Wang, Y.; Tao, H.; He, W.; Lv, G.; and Chen, E. 2020. Context-Aware Generation-Based Net For Multi-Label Visual Emotion Recognition. In *Proceedings of ICME 2020*, 1–6.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2009. The Graph Neural Network Model. *IEEE Trans. Neural Networks* 20(1): 61–80.

Shen, X.; Boutell, M. R.; Luo, J.; and Brown, C. M. 2004. Multilabel machine learning and its application to semantic scene classification. In *Proceedings of SPIESR 2004*, 188–199.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15(1): 1929–1958.

Sun, Z.; Sarma, P. K.; Sethares, W. A.; and Liang, Y. 2020. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. In *Proceedings of AAAI 2020*, 8992–8999.

Tsai, Y. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of ACL 2019*, 6558–6569.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. P. 2011. Random k-Labelsets for Multilabel Classification. *IEEE Trans. Knowl. Data Eng.* 23(7): 1079–1089.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L. 2019. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. In *Proceedings of AAAI 2019*, 7216–7223.

Wu, J.; Xiong, W.; and Wang, W. Y. 2019. Learning to Learn and Predict: A Meta-Learning Approach for Multi-Label Classification. In *Proceedings of EMNLP-IJCNLP 2019*, 4353–4363.

Xiao, L.; Huang, X.; Chen, B.; and Jing, L. 2019. Label-Specific Document Representation for Multi-Label Text Classification. In *Proceedings of EMNLP-IJCNLP 2019*, 466–475. Association for Computational Linguistics.

Yang, P.; Luo, F.; Ma, S.; Lin, J.; and Sun, X. 2019. A Deep Reinforced Sequence-to-Set Model for Multi-Label Classification. In *Proceedings of ACL 2019*, 5252–5258.

Ying, W.; Xiang, R.; and Lu, Q. 2019. Improving Multi-label Emotion Classification by Integrating both General and Domain-specific Knowledge. In *Proceedings of W-NUT@EMNLP 2019*, 316–321.

You, R.; Guo, Z.; Cui, L.; Long, X.; Bao, Y.; and Wen, S. 2020. Cross-Modality Attention with Semantic Graph Embedding for Multi-Label Classification. In *Proceedings of AAAI 2020*, 12709–12716.

Yu, J.; Marujo, L.; Jiang, J.; Karuturi, P.; and Brendel, W. 2018. Improving Multi-label Emotion Classification via Sentiment Classification with Dual Attention Transfer Network. In *Proceedings of EMNLP 2018*, 1097–1102.

Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of ACL 2018*, 2236–2246.

Zhang, D.; Ju, X.; Li, J.; Li, S.; Zhu, Q.; and Zhou, G. 2020a. Multi-modal Multi-label Emotion Detection with Modality and Label Dependence. In *Proceedings of EMNLP 2020*, 3584–3593.

Zhang, D.; Wu, L.; Li, S.; Zhu, Q.; and Zhou, G. 2019a. Multi-Modal Language Analysis with Hierarchical Interaction-Level and Selection-Level Attentions. In *Proceedings of ICME 2019*, 724–729.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019b. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *Proceedings of IJCAI 2019*, 5415–5421.

Zhang, D.; Zhang, W.; Li, S.; Zhu, Q.; and Zhou, G. 2020b. Modeling both Intra- and Inter-modal Influence for Real-Time Emotion Detection in Conversations. In *Proceedings of ACM MM 2020*, 503–511. doi:10.1145/3394171.3413949.