Multi-modal Multi-label Emotion Detection with Modality and Label Dependence

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li*, Qiaoming Zhu, Guodong Zhou
School of Computer Science and Technology, Soochow University, China
dzhang@suda.edu.cn, xcju@stu.suda.edu.cn
{jhli, lishoushan, qmzhu, gdzhou}@suda.edu.cn

Abstract

As an important research issue in the natural language processing community, multi-label emotion detection has been drawing more and more attention in the last few years. However, almost all existing studies focus on one modality (e.g., textual modality). In this paper, we focus on multi-label emotion detection in a multi-modal scenario. In this scenario, we need to consider both the dependence among different labels (label dependence) and the dependence between each predicting label and different modalities (modality dependence). Particularly, we propose a multi-modal sequence-to-set approach to effectively model both kinds of dependence in multi-modal multi-label emotion detection. The detailed evaluation demonstrates the effectiveness of our approach.

1 Introduction

Emotion detection is to predict emotion categories, such as *angry*, *happy*, and *surprise*, expressed by an utterance of a speaker and has largely encompassed a variety of applications, such as online chatting (Galik and Rank, 2012; Zhang et al., c), news analysis (Li et al., 2015; Zhu et al., 2019) and dialogue systems (Ghosal et al., 2019; Zhang et al., d). Over the last few years, there has been a substantial body of research on emotion detection (Abdul-Mageed and Ungar, 2017; Zhou et al., 2019; Zhang et al., a), where a considerable amount of work has focused on multi-label emotion detection (Li et al., 2015; Yu et al., 2018; Ying et al., 2019).

Basically, emotion detection is a multi-label classification problem since one utterance naturally tends to involve more than one emotion category. However, classifying instances with multiple possible categories is sometimes much more difficult than classifying instances with a single label. One main challenge is how to model the **label dependence** in the classification approach. For example, in the utterance as shown in Figure 1, both the *Sad* and *Disgust* emotions are more likely to exist, rather than the conflicting emotions of *Sad* and *Happy*. Recent studies, such as (Yang et al., 2019) and (Xiao et al., 2019), have begun to address this challenge.

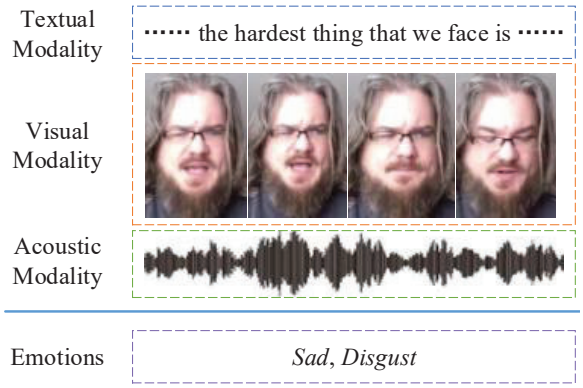

Figure 1: An example of multi-modal instance with multi-label emotion categories in a video segment.

However, almost all existing studies in multi-label emotion detection focus on one modality (e.g., textual modality). Only very recently, the research community has become increasingly aware of the need on multi-modal emotion detection (Zadeh et al., 2018b) due to its wide potential applications, e.g., with the massively growing importance of analyzing conversations in speech (Gu et al., 2019) and video (Majumder et al., 2019). In this study, we aim to tackle multi-modal multi-label emotion detection. Compared with single modality, multi-modal multi-label emotion detection needs to well model the contribution of different modalities for each label since each modality has a different impact on expressing emotion. For example, from the textual modality as shown in Figure 1, while we may only infer the *Sad* emotion, we are more

*Corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3584–3593,
November 16–20, 2020.

likely to infer the *Disgust* emotion instead from the visual modality. Meanwhile, the acoustic modality may not help label prediction in this case. Therefore, besides the label dependence, it is important and challenging to effectively model another dependence, namely the **modality dependence**.

In this paper, we address the above challenges in multi-modal multi-label emotion detection by proposing a multi-modal seq2set (MMS2S) approach to model both the modality and label dependence simultaneously. Specifically, while the conditional generation framework naturally models the label dependence by predicting the next emotion label upon other potential labels, we propose Multi-head soft modality attention at each predicting step inside the emotion decoder to capture the modality dependence. First, we adopt three single-modal encoders based on Transformer to capture the single-modal characteristics of the textual, visual and acoustic modalities, respectively. Then, we make the given emotion representation attend to three intra-modal sequences from encoders inside the emotion decoder and leverage multi-head soft modality attention to control the different contributions of different modalities for each potential emotion prediction. Finally, we train our proposed model by maximizing the probabilities of top K sequences and predict all potential emotion labels by finding the most likely emotion label set.

Systematical evaluation on a public multi-modal multi-label emotion dataset, i.e., CMU-MOSEI, shows that our approach significantly outperforms several state-of-the-art baselines.

2 Related Work

As an interdisciplinary research field, emotion detection has been drawing more and more attention in both natural language processing and multi-modal communication (Zadeh et al., 2018c). In the NLP community, almost all existing studies of multi-label emotion detection rely on special knowledge of emotion, such as context information (Li et al., 2015), cross-domain transferring (Yu et al., 2018) and external resource (Ying et al., 2019). In fact, when there is no special knowledge (Kim et al., 2018), it can be normally handled by multi-label text classification approaches. In the multi-modal community, related studies normally focus on single-label emotion task and the studies for multi-label emotion task are much less and limited to be transformed to multiple binary classification (Zadeh et al., 2018b; Wang et al., 2019; Akhtar et al., 2019; Chauhan et al., 2019). In the following, we give an overview of multi-label emotion/text classification and multi-modal emotion detection.

Multi-label Emotion/Text Classification. Recent studies normally cast multi-label emotion detection task as a classification problem and leverage the special knowledge as auxiliary information (Yu et al., 2018; Ying et al., 2019). These approaches may not be easily extended to those tasks without external knowledge. At this time, the multi-label text classification approaches can be quickly applied to emotion detection. There have been a large number of representative studies for that. Kant et al. (2018) leverage the pre-trained BERT to perform multi-label emotion task and Kim et al. (2018) propose an attention-based classifier that predicts multiple emotions of a given sentence. More recently, Yang et al. (2018) propose a sequence generation model and 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.

Different from above studies, we focus on multi-label emotion detection in a multi-modal scenario by considering the modality dependence besides the label dependence. To the best of our knowledge, this is the first attempt to perform multi-label emotion detection in a multi-modal scenario.

Multi-modal Emotion Detection. Recent studies on multi-modal emotion detection largely depend on multi-modal fusion framework to perform binary classification within each emotion category. Recently, Wang et al. (2019) introduce a recurrent attended variation embedding network for multi-modal language analysis with non-verbal shifted word representation. Tsai et al. (2019) employ the Transformer-based architecture to capture the long-range interactions inside and across different modalities. However, they still cast the multi-label emotion detection as multiple binary classification problems.

Different from above studies, we focus on multi-modal emotion detection in a multi-label scenario by considering the label dependence besides the modality dependence. To the best of our knowledge, this is the first attempt to perform multi-modal emotion detection in a multi-label scenario.

3 Data Pre-processing

We extract low-level handcrafted features from three modalities. First of all, we align the three modalities by extracting the exact utterance timestamp of each word using P2FA (Yuan and Liberman, 2008). Since words are considered as the basic semantic units of language, we use the interval duration of each word as a time-step. Then, we calculate the expected video and audio features by taking the expectation of their feature values over the word time interval (Zadeh et al., 2018a; Zhang et al., b). On this basis, we process the information of the three modalities as follows.

Textual Modality. The GloVe word embeddings (Pennington et al., 2014) are used to represent the words from manual transcripts. Then, we get the text sequence $X^T = [x_1^T, x_2^T, \cdots, x_m^T]$ with dimension $m \times d^T$.

Visual modality. The library Facet [1] is used to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features (Zadeh et al., 2018c) to form a sequence of facial gesture throughout time. Then, we get the visual sequence $X^V = [x_1^V, x_2^V, \cdots, x_m^V]$ with dimension $m \times d^V$.

Acoustic Modality. The COVAREP software (Degottex et al., 2014) is used to extract acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segmenting features (Drugman and Alwan, 2011), glottal source parameters (Drugman et al., 2012), peak slope parameters and maxima dispersion quotients (Kane and Gobl, 2013). Then, we get the acoustic sequence $X^A = [x_1^A, x_2^A, \cdots, x_m^A]$ with dimension $m \times d^A$.

4 Multi-modal Seq2Set for Multi-modal Multi-label Emotion Detection

4.1 Problem Description

In this section, we define some notations and describe the multi-modal multi-label emotion detection (MMED) task. Given the label space with L labels $\mathcal{L} = \{l_1, l_2, \cdots, l_L\}$, the textual, visual and acoustic sequences, i.e., X^T, X^V and X^A containing m time steps respectively, the task is to assign a subset $\boldsymbol{y}$ containing L' labels in the label space $\mathcal{L}$, i.e., $\{y_1, y_2, \cdots, y_{L'}\}$. Unlike traditional single-label classification where only one label is assigned to each sample, each sample in the MMED task can have multiple labels. In previous studies (Yang et al., 2018, 2019), from the perspective of sequence generation, the MMED task can be modeled as finding an optimal label sequence $\boldsymbol{y}^*$ that maximizes the conditional probability $p(\boldsymbol{y}^*|X^T, X^V, X^A)$. Although sequence decoding by conditioned on previous steps can effectively capture the dependence among an output sequence, all possible emotion labels of an utterance are a set rather than a fixed sequence. Therefore, we adopt a conditional set generation mechanism, which maximizes the log-likelihood as follows:

$$\sum_{i=1}^{Num} log \sum_{\mathrm{s} \in \pi(\boldsymbol{y}^i)} p(\mathrm{s}|(X^T)^i, (X^V)^i, (X^A)^i) \quad (1)$$

where Num denotes the total number of multi-modal samples in the dataset. $\pi(\boldsymbol{y}^i)$ stands for all permutations of the label set $\boldsymbol{y}^i$ of the i-th sample.

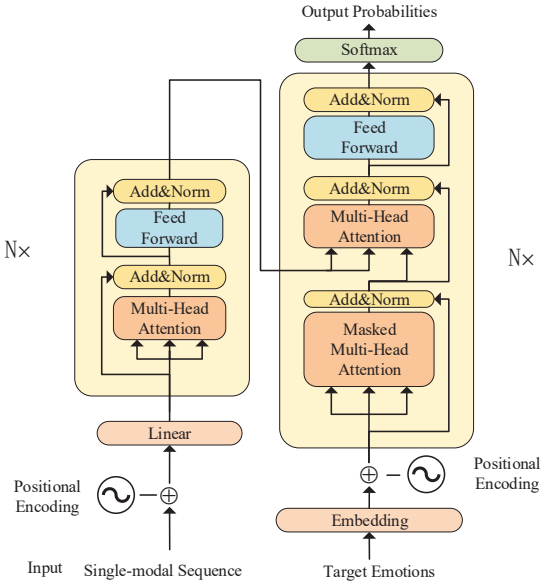

Figure 2: The architecture of a standard Transformer (Vaswani et al., 2017) for sequence to sequence learning.

4.2 Background

Since our approach is based on Transformer architecture, we give a brief description of a standard Transformer (Vaswani et al., 2017) for seq2seq learning as shown in Figure 2.

The encoder is composed of a stack of N identical layers, each of which has two sub-layers. The first sub-layer is a multi-head self-attention network, and the second one is a position-wise fully connected feed-forward network. A residual connection (He et al., 2016) is employed around each

[1] https://imotions.com/emotient/

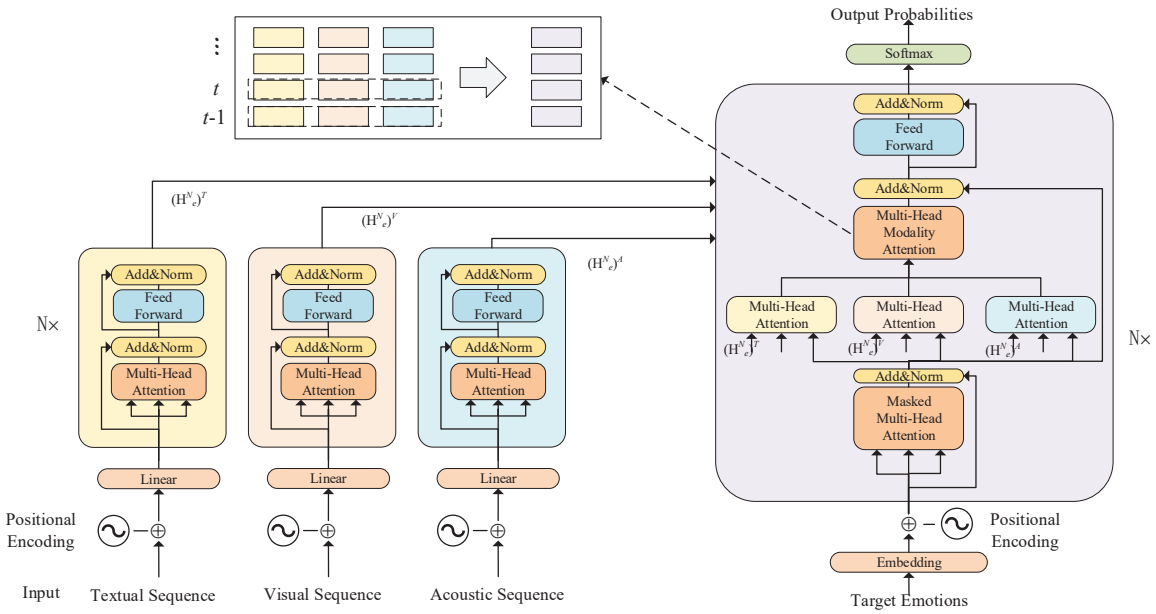

Figure 3: The overview of multi-modal seq2set.

sub-layer, followed by layer normalization (Ba et al., 2016). Formally, the output of the first sub-layer C_e^n and the second sub-layer H_e^n at n-th layer are sequentially calculated as:

$$\mathrm{C}_e^n = \mathrm{LN}(\mathrm{SATT}(\mathrm{H}_e^{n-1}) + \mathrm{H}_e^{n-1}) \quad (2)$$

$$\mathrm{H}_e^n = \mathrm{LN}(\mathrm{FFN}(\mathrm{C}_e^n) + \mathrm{C}_e^n) \quad (3)$$

where $\mathrm{LN}(\cdot)$, $\mathrm{SATT}(\cdot)$, and $\mathrm{FFN}(\cdot)$ are layer normalization, multi-head self-attention mechanism, and feed-forward network with ReLU activation, respectively. The subscript e denotes the encoding part.

The decoder is also composed of a stack of N identical layers. In addition to two sub-layers in each decoder layer, the decoder inserts a third sub-layer D_d^n to perform attention over the output of the encoder H_e^N:

$$\mathrm{C}_d^n = \mathrm{LN}(\mathrm{SATT}(\mathrm{H}_d^{n-1}) + \mathrm{H}_d^{n-1}) \quad (4)$$

$$\mathrm{D}_d^n = \mathrm{LN}(\mathrm{ATT}(\mathrm{C}_d^n, \mathrm{H}_e^N) + \mathrm{C}_d^n) \quad (5)$$

$$\mathrm{H}_d^n = \mathrm{LN}(\mathrm{FFN}(\mathrm{C}_d^n) + \mathrm{C}_d^n) \quad (6)$$

where $\mathrm{ATT}(\mathrm{C}_d^n, \mathrm{H}_e^N)$ denotes attending the top output of encoder H_e^N with C_d^n as query. The subscript d denotes the decoding part. The top layer output of the decoder H_d^N is used to generate the final output sequence.

4.3 Multi-modal Seq2Set Approach

Figure 2 shows the overall architecture of our proposed Multi-Modal Seq2Set (MMS2S) approach. Note that the novel decoding module can well handle the modality and label dependence by soft modality attention and conditional label generation.

Multi-modal Sequences Encoding. We first build three independent Transformer-based encoders to capture the temporal information and self-modal dynamics in each modality. Formally,

$$(\mathrm{H}_e^N)^M = \mathrm{Trans}_e^M(X^M) \quad (7)$$

where $M \in \{T, V, A\}$ denotes the symbol of modality. $(\mathrm{H}_e^N)^M \in \mathbb{R}^{m \times d_m}$ denotes the final output of the encoder for modality M and Trans_e^M denotes a Transformer-based encoder function for modality M.

From the multi-modal sequences encoding module, we can obtain the feature sequence of each modality: $(\mathrm{H}_e^N)^T$, $(\mathrm{H}_e^N)^V$ and $(\mathrm{H}_e^N)^A$.

Multi-Head Modality Attention. All modality-specific sequences are simultaneously fed into the decoding module. For each decoding step, the decoder attends to the encoding representation of each modality independently. Formally,

$$\mathrm{C}_d^n = \mathrm{LN}(\mathrm{SATT}(\mathrm{H}_d^{n-1}) + \mathrm{H}_d^{n-1}) \quad (8)$$

$$(\mathrm{C}_{d\to e}^n)^M = \mathrm{ATT}(\mathrm{C}_d^n, (\mathrm{H}_e^N)^M) \quad (9)$$

Then, we can obtain three contextual sequences from decoder attending to encoders: $(\mathrm{C}_{d\to e}^n)^T$, $(\mathrm{C}_{d\to e}^n)^V$ and $(\mathrm{C}_{d\to e}^n)^A$. We leverage multi-head modality attention over three sequences to control different contribution of different modalities at each step for feature matrix $(\mathrm{C}_{d\to e}^n)_t =$

Algorithm 1 Training Procedure for MMS2S

Input: Multi-modal Multi-label Dataset $(X^i, \boldsymbol{y}^i)$, $X^i = ((X^T)^i, (X^V)^i, (X^T)^i)$, $i = 1, 2, \cdots, Num$;
Output: Trained parameters of MMS2S model ;

```
for each batch do
    for each (X^i, y^i) in a batch do
        Get top K sequences by beam search and their probabilities:
        {s_1^i, ..., s_K^i; p(s_1^i|X^i), ..., p(s_K^i|X^i)};
    end for
    Update model parameters by maximizing:
    sum_{(X^i, y^i) in batch} log sum_{s in {s_1^i, ..., s_K^i}} p(s|X^i)
end for
```

$[(\mathrm{C}^n_{d\to e})^T_t, (\mathrm{C}^n_{d\to e})^V_t, (\mathrm{C}^n_{d\to e})^A_t] \in \mathbb{R}^{3\times d_m}$. Formally,

$$\mathrm{C}_t = \mathrm{SATT}((\mathrm{C}^n_{d\to e})_t) \quad (10)$$

$$(\mathrm{C}^n_d)'_t = W_s(\mathrm{C}^T_t \oplus \mathrm{C}^V_t \oplus \mathrm{C}^A_t) \quad (11)$$

where $\mathrm{C}_t = [\mathrm{C}^T_t, \mathrm{C}^V_t, \mathrm{C}^A_t] \in \mathbb{R}^{3\times d_m}$ denotes the temporary multi-modal hybrid representation at t-th step. $(\mathrm{C}^n_d)'_t \in \mathbb{R}^d_m$ denotes the feature vector by soft modality weighting the t-th step. $W_s \in \mathbb{R}^{d_m\times 3d_m}$ is a trainable matrix to scale the dimension of multi-modal hybrid representation. $\oplus$ denotes the concatenating operation.

Subsequently, as the normal propagation, the adaptive contextual sequence is fed into the feed-forward layer,

$$\mathrm{D}^n_d = \mathrm{LN}((\mathrm{C}^n_d)' + \mathrm{C}^n_d) \quad (12)$$

$$\mathrm{H}^n_d = \mathrm{LN}(\mathrm{FFN}(\mathrm{C}^n_d) + \mathrm{C}^n_d) \quad (13)$$

Emotion Prediction. Finally, the top output of decoder $\mathrm{Z} = \mathrm{H}^N_d \in \mathbb{R}^{m'\times d_m}$ is used to predict all potential emotions via linear and softmax layer. Formally,

$$p_t = \mathrm{softmax}(\mathrm{Z}_t W_p + I_t) \quad (14)$$

where $W_p \in \mathbb{R}^{d_m\times L}$, is a trainable weight matrix. $I_t \in \mathbb{R}^L$ is the mask vector that is used to prevent the decoder from predicting repeated labels:

$$(I_t)_k = \begin{cases} -\infty & \text{if label } l_k \text{ has been predicted} \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

Training by Top K Sequences. We approximate the objective of Eq. 1 by only considering the top K highest probability sequences produced by our model. We leverage a variant of beam search (Qin et al., 2019) for sets with width K. In particular, the search candidates in each step are restricted to only labels in the golden set. This approximates inference procedure is carried out repeatedly before each batch training step to find highest probability sequences for all training instances occurring in that batch. Algorithm 1 shows the detailed procedure.

Algorithm 2 Testing Procedure for MMS2S

Input: Multi-modal Instance X, $X = (X^T, X^V, X^T)$;
Output: Predicted Emotion Label Set $\hat{\boldsymbol{y}}$;

```
Obtain K highest probability sequences by beam search: {s_1, ..., s_K};
Map each sequence s_k to the corresponding set y_k and remove duplicate sets (if any);
for each y_k do
    Get top K sequences associated with y_k and their probabilities by beam search:
    {s'_1, ..., s'_K; p(s'_1|X), ..., p(s'_K|X)};
    Set probability is approx. by summing up:
    p(y_k|X) ≈ sum_{s in {s'_1, ..., s'_K}} p(s|X);
end for
ŷ = argmax_{y_k} p(y_k|X)
```

Testing by Most Probable Set. Different from the previous approach of directly using most probable sequence as a set based on a pre-defined baseline (Yang et al., 2019) , we instead aim to find the most likely set, which involves summing up probabilities for all its permutations. Algorithm 2 shows the detailed procedure. Note that both the training and testing procedures allow our model to be far more freedom on label order.

5 Experimentation

In this section, we systematically evaluate our approach to multi-modal multi-label emotion detection.

5.1 Experimental Settings

Dataset. We use the largest available multi-modal emotion benchmark dataset, i.e., CMU-MOSEI (Zadeh et al., 2018b) in our evaluation. The dataset is 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*. The average words of utterance-level video clips are

Multi-label	Number	Emotion	Number
None	3372	*Happiness*	12240
One	11050	*Surprise*	1892
Two	5526	*Sadness*	5918
Three	2084	*Anger*	4933
Four	553	*Disgust*	3680
Five	84	*Fear*	2286
Six	8	-	-

Table 1: The statistics on the CMU-MOSEI dataset.

19.1 and the average number of emotion labels per sample is 1.6. The training, validation and test data are all the same with the split videos and utterances available in the public SDK[2]. Table 1 shows the brief statistics of the samples with multiple labels.

Implementation Details. We implement our approach via Pytorch toolkit (torch-0.4.1) with a piece of GTX 1080 Ti. Following (Zadeh et al., 2018b), the textual input dimension d^T is set to 300, the visual input dimension d^V is set to 35 and the acoustic input dimension d^A is set to 74. The hidden size d_m in the encoders and decoder is 512. The number of heads in SATT and ATT is 8.

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. We adopt cross-entropy as the loss function and use the Adam (Kingma and Ba, 2014) optimization method to minimize the loss over the training data. For the hyper-parameters of the Adam optimizer, we set the learning rate as 0.001 with two momentum parameters of β_1 and β_2, 0.9 and 0.999 respectively. The beam size K is set to be 5 at both training and inference stages. To motivate future research, the code will be released via github [3].

Evaluation Metrics and Significance Test. In our study, we employ three evaluation metrics to measure the performances of different approaches to multi-modal multi-label emotion detection, i.e., multi-label Accuracy (*Acc*), Hamming Loss (*HL*) and micro F_1 measure (F_1). These metrics have been popularly used in some multi-label classification problems (Li et al., 2015; Yang et al., 2019; Aly et al., 2019; Wu et al., 2019).

Note that smaller Hamming Loss corresponds to better classification quality, while larger Accuracy and F_1 measure corresponds to better classification quality. Besides, through scipy[4], the paired t-test is performed to test the significance of the difference between two approaches, with a default significant level of 0.05.

5.2 Baselines

For a thorough comparison, we implement various baseline approaches in three groups:

Multi-label Classification Approaches. In this group, the baselines use different approaches to deal with the multi-label issue without considering the modality dependence issue. Specifically, in these approaches, the multi-modal inputs are early fused (simply concatenated) as a new input. (1) **BR**[5] (Shen et al., 2004), which transforms the multi-label task into multiple single-label binary classification problems by ignoring the correlations between labels. (2) **CC**[5] (Read et al., 2011), which transforms the multi-label task into a chain of binary classification problems and takes high-order label correlations into consideration. (3) **RAkEL**[5] (Tsoumakas et al., 2011), which improves the Label Powerset (Tsoumakas and Katakis, 2007) with breaking the initial set of labels into a number of small random subsets and training a corresponding classifier. (4) **AC**[6] (Kim et al., 2018), which consists of a self-attention module and multiple CNNs enabling it to imitate human's two-step procedure of analyzing emotions from sentences: comprehend and classify. (5) **LSAN**[7] (Xiao et al., 2019), which takes advantage of label semantic information to determine the semantic connection between labels and document for constructing label-specific document representation. This approach is considered as the state-of-the-art in multi-label text classification. (6) **DRS2S**[8] (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. This approach is also considered as the state-of-the-art in multi-label text classification.

Multi-modal Classification Approaches. In this group, the baselines use different approaches to deal with the multi-modal issue without considering the label dependence issue. Specifically, in these approaches, a linear layer of L dimensions

[2]https://github.com/A2Zadeh/CMU-MultimodalSDK
[3]https://github.com/MANLP-suda/MMS2S
[4]https://www.scipy.org/
[5]http://scikit.ml/
[6]https://github.com/yanghoonkim/attnconvnet
[7]https://github.com/EMNLP2019LSAN/LSAN/
[8]https://github.com/lancopku/Seq2Set

Approaches	*Acc*	*HL*	F_1
BR (Shen et al., 2004)	0.222	0.371	0.386
CC (Read et al., 2011)	0.225	0.377	0.386
RAkLA (Tsoumakas et al., 2011)	0.242	0.376	0.397
AC (Kim et al., 2018)	0.388	0.240	0.492
LSAN (Xiao et al., 2019)	0.393	0.209	0.501
DRS2S (Yang et al., 2019)	0.436	0.215	0.523
GMFN (Zadeh et al., 2018b)	0.396	0.195	0.517
RAVEN (Wang et al., 2019)	0.416	0.195	0.517
MulT (Tsai et al., 2019)	0.445	0.190	0.531
MMS2S (Ours)	**0.475**	**0.182**	**0.560**
MMS2S w/o M	0.421	0.225	0.525
MMS2S w/o L	0.417	0.212	0.523

Table 2: Performance of different approaches to multi-modal multi-label emotion detection.

with $sigmoid$ activation is used to predict the emotions. (7) **GMFN**[2] (Zadeh et al., 2018b), which explicitly models the multi-modal interactions by capturing uni-modal, bi-modal and tri-modal interactions. (8) **RAVEN**[9] (Wang et al., 2019), which models the fine-grained structure of nonverbal subword sequences and dynamically shifts word representations based on nonverbal cues. This approach is considered as the state-of-the-art in multi-modal language analysis. (9) **MulT**[10] (Tsai et al., 2019), which addresses long-range dependencies between elements across modalities in an end-to-end manner. This approach is considered as the state-of-the-art in multi-modal emotion detection.

Ablated Approaches. (10) **MMS2S w/o M**, a variation of our approach, which replaces the multi-head modality attention with simply concatenation. (11) **MMS2S w/o L**, a variation of our approach, which replaces the decoder with $sigmoid$ activation for L dimension.

5.3 Experimental Results

Comparison with the multi-modal and multi-label classification approaches. Table 2 shows the results of different approaches to multi-modal multi-label emotion detection. From this table, we can see that (1) the classical multi-label approaches **BR**, **CC** and **RAkLA** perform much worse than the deep learning baselines **AC**, **LSAN** and **DRS2S**. For instance, **DRS2S** outperforms **RAkLA** by 19.4%, 16.1% and 12.6% with respect to *Acc*, *HL* and F_1, respectively. This indicates that the approaches with deep representation do have more advantages than the classical approaches towards multi-label problem. (2) The baselines of multi-modal classification outperform the baselines of text-based multi-label classification in most cases. Especially, **MulT** performs much better than **LSAN** and **DRS2S** in terms of all metrics. This is mainly due to the fact that multi-modal data need to well model the intra-modal and inter-modal dynamics and the early fusion approaches inevitably result in performance loss. (3) Among all the approaches, our proposed **MMS2S** performs best in terms of all metrics. The t-test demonstrates that our approach significantly outperforms **LSAN**, **DRS2S**, , **a**nd **MulT**, respectively (p-value < 0.05).

[9]https://github.com/victorywys/RAVEN

[10]https://github.com/yaohungt/Multimodal-Transformer

Num.	Approaches	*Acc*	*HL*	F_1
1	DRS2S	0.415⇓	0.242⇓	0.514⇓
	MMS2S (Ours)	**0.475-**	**0.183**↓	**0.560-**
2	DRS2S	0.419⇓	0.227⇓	0.506⇓
	MMS2S (Ours)	**0.473**↓	**0.185**↓	**0.559**↓

Table 3: The impact of random label order as ground-truth. ⇓: Significant decrease, ↓: Insignificant decrease, −: No decrease.

Ablation Study. To further demonstrate the importance of modeling modality and label dependence, we do not model either the modality (**MMS2S w/o M**) or the label dependence (**MMS2S w/o L**). From Table 2, we observe that not modeling either the modality or the label dependency significantly decreases the performance. This illustrates the effectiveness of our approach in modeling the two types of dependence.

5.4 Analysis

Impact of random label order. We investigate the impact of random label order as ground-truth for our proposed **MMS2S** and a previous text-based seq2set approach **DRS2S** (Yang et al., 2019) as shown in Table 3. We can observe that **DRS2S** largely depends on the pre-defined knowledge of label order (such as descending order of label frequency). While, our approach performs well with random label order as ground-truth, suggesting that our approach indeed generates an adaptive emotion label set rather than a sequence for each sample.

Single-modal approach vs. multi-modal approach. To illustrate the necessity of multi-modal approach for multi-label emotion detection, we also evaluate single-modal approach via sequence-to-set training and testing, namely **SMS2S**, which aims at modeling a single modality while ignoring the other two. Table 4 compares the performance of SMS2S and MMS2S approaches. From this ta-

Approach	Modality	*Accuracy*	*Hamming Loss*	*F_1 measure*	*Precision*	*Recall*
SMS2S	Text	0.438	0.216	0.492	0.561	0.438
	Vision	0.396	0.221	0.440	0.505	0.390
	Audio	0.395	0.219	0.451	0.552	0.381
MMS2S	Text&Vision&Audio	**0.475**	**0.182**	**0.560**	**0.576**	**0.545**

Table 4: Performance of single-modal and multi-modal seq2set approaches.

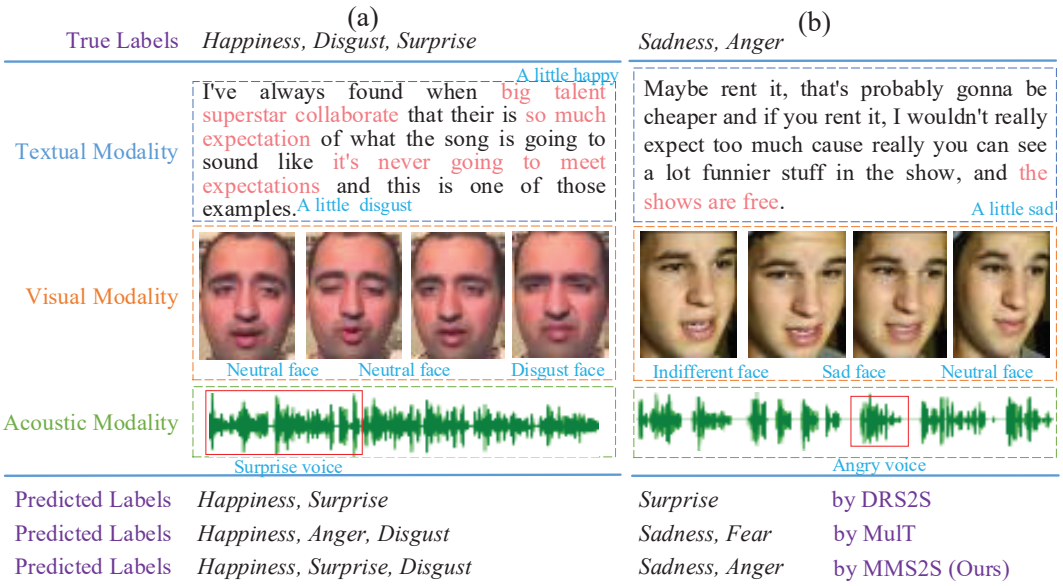

Figure 4: Two cases of the predicted labels by **DRS2S**, **MulT** and **MMS2S**.

ble, we observe that **SMS2S** with textual modality outperforms the counterparts with the other two modalities, suggesting that the textual modality contains more useful information than the others. Moreover, our **MMS2S** achieves the highest performance, suggesting that both the visual and the acoustic modalities could be useful complement to the textual modality. This is consistent with our motivation that different modality plays different roles in emotion expressing.

Case Study. To further demonstrate the effectiveness of our multi-modal seq2set approach, Figure 4 presents two examples with predicted emotions by **MMS2S**, and two representative baselines **DRS2S** and **MulT**. We take the case (a) as an example: although **DRS2S** can accurately detect two emotions of the ground-truth, it leaves the *Disgust* emotion. This is mainly because early fusion without modality dependence results in different modalities information confusion so that it may be difficult for **DRS2S** to infer all the correct emotions. In contrast, **MulT** can detect the *Disgust* emotion and obtain three emotions. But it gives a wrong prediction of *Anger*. Obviously, *Happiness* and *Anger* are conflicting emotions. This indicates that **MulT** completely ignores the label dependence. However, from both cases, we observe that our **MMS2S** can obtain all correct emotions by properly modeling modality dependence and label dependence.

6 Conclusion

In this paper, we propose a multi-modal sequence-to-set approach to simultaneously handle the modality and label dependence in multi-modal multi-label emotion detection. Our approach can not only model the dependence between each label and different modalities, but also model the dependence among multiple labels of a sample. The detailed evaluation demonstrates that our proposed model significantly outperforms several state-of-the-art baselines.

In our future work, we will extend our approach to more multi-modal multi-label scenarios, such as intention detection in video conversations and aspect analysis in multi-modal reviews. Furthermore, we would like to investigate other approaches (e.g., graph-based neural network) to better model the modality and label dependence in multi-modal multi-label emotion detection.

Acknowledgments

We thank all anonymous reviewers for their helpful comments. This work was supported by three NSFC grants, i.e., No. 61672366, No. 61876120 and No. 61836007. This work was also supported by a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of ACL 2017*, pages 718–728.

Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of NAACL-HLT 2019*, pages 370–379.

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of ACL 2019*, pages 323–330.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of EMNLP 2019*, pages 5651–5661.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *Proceedings of IEEE ICASSP 2014*, pages 960–964.

Thomas Drugman and Abeer Alwan. 2011. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of INTERSPEECH 2011*, pages 1973–1976.

Thomas Drugman, Mark R. P. Thomas, Jón Gunason, Patrick A. Naylor, and Thierry Dutoit. 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE TASLP*, 20(3):994–1006.

Maros Galik and Stefan Rank. 2012. Modelling emotional trajectories of individuals in an online chat. In *Proceedings of MATES 2012*, pages 96–105.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of EMNLP 2019*, pages 154–164.

Yue Gu, Xinyu Lyu, Weijia Sun, Weitian Li, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2019. Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition. In *Proceedings of ACM MM 2019*, pages 157–166.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR 2016*, pages 770–778.

John Kane and Christer Gobl. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE TASLP*, 21(6):1170–1179.

Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Yanghoon Kim, Hwanhee Lee, and Kyomin Jung. 2018. Attnconvnet at semeval-2018 task 1: Attention-based convolutional neural networks for multi-label emotion classification. In *Proceedings of SemEval@NAACL-HLT 2018*, pages 141–145.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level emotion classification with label and context dependence. In *Proceedings of ACL 2015*, pages 1045–1053.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proceedings of AAAI 2019*, pages 6818–6825.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

Kechen Qin, Cheng Li, Virgil Pavlu, and Javed A. Aslam. 2019. Adapting RNN sequence prediction model to multi-label set prediction. In *Proceedings of NAACL-HLT 2019*, pages 3181–3190.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Xipeng Shen, Matthew R. Boutell, Jiebo Luo, and Christopher M. Brown. 2004. Multilabel machine learning and its application to semantic scene classification. In *Proceedings of SPIESR 2004*, pages 188–199.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL 2019*, pages 6558–6569.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *IJDWM*, 3(3):1–13.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. 2011. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.*, 23(7):1079–1089.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS 2017*, pages 5998–6008.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of AAAI 2019*, pages 7216–7223.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. In *Proceedings of EMNLP 2019*, pages 4345–4355.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of EMNLP 2019*, pages 466–475.

Pengcheng Yang, Fuli Luo, Shuming Ma, Junyang Lin, and Xu Sun. 2019. A deep reinforced sequence-to-set model for multi-label classification. In *Proceedings of ACL 2019*, pages 5252–5258.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. SGM: sequence generation model for multi-label classification. In *Proceedings of COLING 2018*, pages 3915–3926.

Wenhao Ying, Rong Xiang, and Qin Lu. 2019. Improving multi-label emotion classification by integrating both general and domain-specific knowledge. In *Proceedings of W-NUT 2019*, pages 316–321.

Jianfei Yu, Luís Marujo, Jing Jiang, Pradeep Karuturi, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. In *Proceedings of EMNLP 2018*, pages 1097–1102.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(123):3878.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of AAAI 2018*, pages 5634–5641.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of ACL 2018*, pages 2236–2246.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018c. Multi-attention recurrent network for human communication comprehension. In *Proceedings of AAAI 2018*, pages 5642–5649.

Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. a. Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language. In *Proceedings of ACM MM 2019*, pages 148–156.

Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. b. Modeling the clause-level structure to multimodal sentiment analysis via reinforcement learning. In *Proceedings of IEEE ICME 2019*, pages 730–735.

Dong Zhang, Liangqing Wu, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. c. Multi-modal language analysis with hierarchical interaction-level and selection-level attentions. In *Proceedings of IEEE ICME 2019*, pages 724–729.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. d. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of IJCAI 2019*, pages 5415–5421.

Xiabing Zhou, Zhongqing Wang, Shoushan Li, Guodong Zhou, and Min Zhang. 2019. Emotion detection with neural personal discrimination. In *Proceedings of EMNLP 2019*, pages 5502–5510.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of ACL 2019*, pages 471–480.