Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection

Xincheng Ju
School of Computer Science and Technology, Soochow University, China
xcju@stu.suda.edu.cn

Dong Zhang*
School of Computer Science and Technology, Soochow University, China
dzhang@suda.edu.cn

Junhui Li
School of Computer Science and Technology, Soochow University, China
lijunhui@suda.edu.cn

Guodong Zhou
School of Computer Science and Technology, Soochow University, China
gdzhou@suda.edu.cn

ABSTRACT

Multi-modal utterance-level emotion detection has been a hot research topic in both multi-modal analysis and natural language processing communities. Different from traditional single-label multi-modal sentiment analysis, typical multi-modal emotion detection is naturally a multi-label problem where an utterance often contains multiple emotions. Existing studies normally focus on multi-modal fusion only and transform multi-label emotion classification into multiple binary classification problem independently. As a result, existing studies largely ignore two kinds of important dependency information: (1) Modality-to-label dependency, where different emotions can be inferred from different modalities, that is, different modalities contribute differently to each potential emotion. (2) Label-to-label dependency, where some emotions are more likely to coexist than those conflicting emotions. To simultaneously model above two kinds of dependency, we propose a unified approach, namely multi-modal emotion set generation network (MESGN) to generate an emotion set for an utterance. Specifically, we first employ a cross-modal transformer encoder to capture cross-modal interactions among different modalities, and a standard transformer encoder to capture temporal information for each modality-specific sequence given previous interactions. Then, we design a transformer-based discriminative decoding module equipped with modality-to-label attention to handle the modality-to-label dependency. In the meanwhile, we employ a reinforced decoding algorithm with self-critic learning to handle the label-to-label dependency. Finally, we validate the proposed MESGN architecture on a word-level aligned and unaligned multi-modal dataset. Detailed experimentation shows that our proposed MESGN architecture can effectively improve the performance of multi-modal multi-label emotion detection.

*Corresponding author

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413577

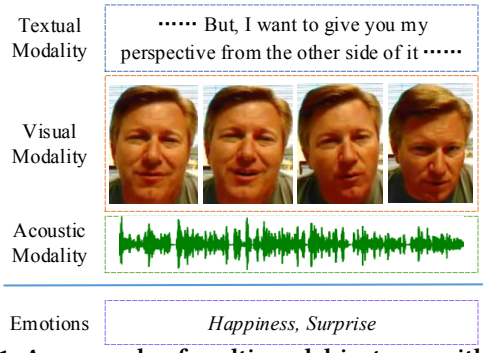

Figure 1: An example of multi-modal instance with multi-label emotion categories.

CCS CONCEPTS

• **Information systems** → **Sentiment analysis**; *Multimedia and multimodal retrieval*; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

multi-modal; multi-label; emotion detection; transformer-based; label set generation

ACM Reference Format:
Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based Label Set Generation for Multi-modal Multi-label Emotion Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413577

1 INTRODUCTION

Predicting emotion categories, such as *angry*, *happy*, and *surprise*, expressed by an utterance of a speaker, encompasses a variety of emerging applications, such as online chatting [4], news analysis [11] and dialogue systems [5, 36]. Over the past decade, there has been a substantial body of research on emotion detection, where a considerable amount of work has focused on text-based [1, 37] emotion classification.

Recently, the research community has become increasingly aware of the huge demand for multi-modal language emotion detection [31, 33] due to its wide potential applications, e.g. the massively growing importance of analyzing conversations in speech [7, 34]

or video [12, 35]. Although some studies [19, 22, 31] turn to the multi-modal scenario and attend to handle multi-label emotion detection, this task still exhibits at least two challenges:

On the one hand, modality-to-label dependency makes multi-modal emotion detection much more difficult to infer multiple labels than uni-modal emotion detection from either text, audio, or vision information. It is common that in an utterance, different modalities may imply different emotions. Such modality-to-label dependency means that predicting each potential label largely depends on different contributions of different modalities. As shown in Figure 1, we may only infer *Happiness* emotion from the visual modality while we opt to get *Surprise* emotion from the combination of both the textual and visual modalities. Meanwhile, the acoustic modality in this example may not help in emotion detection. Therefore, an ideal solution should adaptively fuse the multiple modalities before predicting each potential emotion, respectively.

On the other hand, label-to-label dependency makes conventional multi-modal approaches [2, 19, 30–32], which view multi-label emotion detection as a multiple binary classification task, fail to capture the correlations among emotion labels. Such label-to-label dependency means that predicting each potential label may depend on other potential emotion labels. As shown in Figure 1, *Happiness* and *Surprise* emotions are more likely to co-occur than those conflicting emotion pairs, such as *Happiness* and *Sadness* emotions. Therefore, an ideal solution should be capable of modeling the correlations among the set of emotion labels.

In this paper, we address the above two challenges in multi-modal multi-label emotion detection by simultaneously modeling both the modality-to-label and label-to-label dependencies. Specifically, we propose a unified multi-modal emotion set generation network (MESGN) architecture. First, we build various cross-modal transformer encoders from all three modality channels to capture the cross-modal interactions among textual, visual and acoustic modalities. Then, we feed the modality-specific interactive sequence into three standard transformer encoders to capture temporal information, respectively. Finally, to model the modality-to-label dependency, we design a discriminative decoding module to independently generate the emotion hidden representation and utilize the modality-to-label attention to control the different contributions of three modalities for each potential emotion. In the meanwhile, to model the label-to-label dependency, we design a reinforced decoding block to predict an emotion set by self-critic learning, instead of a pre-defined sequence.

We systematically evaluate our approach on a public multi-modal multi-label emotion dataset, i.e. CMU-MOSEI. Detailed experimentation shows that our proposed approach significantly outperforms the state-of-the-art in various settings of multi-modal emotion detection.

2 RELATED WORK

As an interdisciplinary research field, utterance-level emotion detection has been drawing more and more attention in both natural language processing and multi-modal communication [32]. In the NLP community, almost all existing studies of multi-label emotion detection rely on special knowledge of emotion, such as context information [11], cross-domain [29], and external resource [28]. In fact, when there is no special knowledge [9], it can be normally handled by multi-label text classification approaches [29]. In the multi-modal community, related studies normally focus on single-label emotion tasks and the studies for multi-label emotion tasks are much less and limited to be transformed into multiple binary classifications [2, 22, 31]. In the following, we give an overview of multi-label emotion/text classification and multi-modal emotion detection.

2.1 Multi-label Emotion Detection and Text Classification

Recent studies normally cast a multi-label emotion detection task as a classification problem and leverage the special knowledge as auxiliary information [28, 29]. 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. [8] leverage the pre-trained BERT to perform multi-label emotion tasks and Kim et al. [9] propose an attention-based classifier that predicts multiple emotions of a given sentence. More recently, Yang et al. [27] propose a sequence generation model with a novel decoder structure to solve the multi-label text classification task. Besides, Xiao et al. [25] take advantage of label semantic information to determine the semantic connection between labels and documents for constructing label-specific document representation. Although these approaches perform well in uni-modal (text-based) emotion detection tasks, they cannot be directly extended to the multi-modal scenario.

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

2.2 Multi-modal Emotion Detection

Recent studies on multi-modal emotion detection largely depend on multi-modal fusion frameworks to perform binary classification within each emotion category. Zadeh et al. [31] first propose a graph dynamic fusion network to explicitly account for the uni- and multi-modal interactions, then continuously model them through time, finally adopt a binary classifier for multi-modal emotion detection. Recently, Wang et al. [22] introduce a recurrent attended variation embedding network for multi-modal language analysis with nonverbal shifted word representation. More recently, Tsai et al. [19] introduce the multimodal transformer encoding to perform directional pairwise cross-modal attention, which attends to interactions between multi-modal sequences across distinct time steps and latently adapts streams from one modality to another. Although this approach has a similar encoding network as our study, it only implicitly bridges the correlation between the multiple modalities and a binary emotion label. Note that above all approaches completely ignores the label-to-label dependency.

Different from the above studies, we focus on multi-modal emotion detection in a multi-label scenario by considering the label-to-label dependency besides the modality-to-label dependency. To the

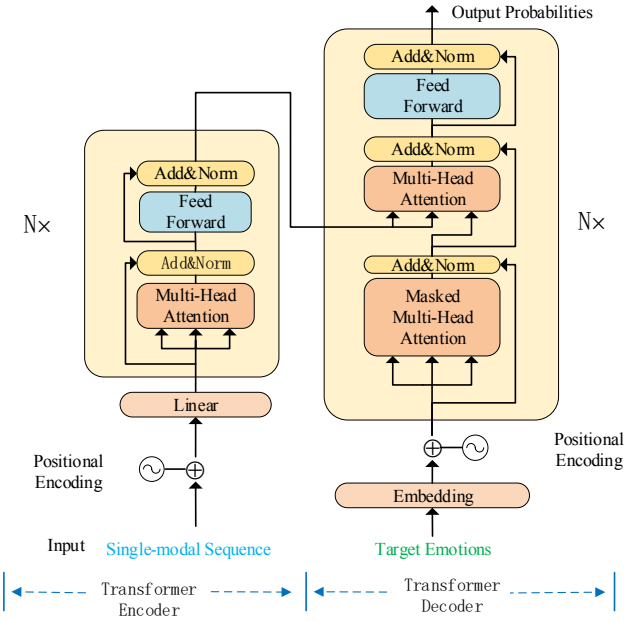

Figure 2: Single-modal Emotion Set Generation Network based on a Transformer

best of our knowledge, this is the first attempt to conduct the research on multi-modal emotion detection in a multi-label scenario.

3 METHODOLOGY

In this section, we describe our methodology as three steps: task definition, traditional single-modal approach, and our proposed multi-modal approach.

3.1 Task Definition

We define some notations and describe the multi-modal multi-label emotion detection (MMED) task. Given the label space with L emotion labels $\mathcal{L} = \{emo_1, emo_2, \cdots, emo_L\}$, and the multi-modal sequences, i.e., Textual sequence X^{Te}, Visual sequence X^{Vi} and Acoustic sequence X^{Ac} containing the words/frames of length n^{Te}, n^{Vi} and n^{Ac} respectively, the task aims 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. From the perspective of the generation task, MMED can be formalized as maximizing the likelihood of given label sets, $\prod^{Num} \sum_{seq \in \pi(\boldsymbol{y})} p(seq|X^{Te}, X^{Vi}, X^{Ac})$, equivalently,

$$\sum_{}^{Num} log \sum_{seq \in \pi(\boldsymbol{y})} p(seq|X^{Te}, X^{Vi}, X^{Ac}) \tag{1}$$

where Num denotes the total number of all training samples. $\pi(\boldsymbol{y})$ denotes all permutations of the emotion set $\boldsymbol{y}$.

3.2 Single-modal Emotion Set Generation

In this section, to better understand our full model MESGN, we introduce the process of single-modal emotion set generation based on prior knowledge of the label order [27]. Here we sort labels by frequency in descending order (from frequent to rare labels). Since transformer-based sequence modeling has been the mainstream architecture, we employ a standard transformer [21] to build the single-modal emotion set generation network (SESGN), as shown in Figure 2. Note that the conditional decoding mechanism can well model the label-to-label dependency naturally.

Given a single-modal sequence $X = [x_1, x_2, \cdots, x_n]$, the encoder maps its corresponding symbol representations $E = [e_1, e_2, \cdots, e_n]$ to a sequence of continuous representations $Z = [z_1, z_2, \cdots, z_n]$, where n is the length of sequence. The decoder then decodes Z into continuous representations $S = [s_1, s_2, \cdots, s_{L'}]$ and generates a emotion set $y = [y_1, y_2, \cdots, y_{L'}]$ one label a time-step. E is the sum of feature embedding and position embedding, where $e_i \in \mathbb{R}^d$.

3.2.1 Single-modal sequence Encoding. The encoder as shown in the left part of Figure 2 is composed of a stack of N identical layers, and each layer has two sub-layers. The first is the self-attention sub-layer and the second is the feed-forward sub-layer. The residual connection is employed around each of the two sublayers, followed by layer normalization. The details of each sub-layer are presented as follows.

$$C^l_{enc} = LN(ATT_{self}(Z^{l-1}_{enc}) + Z^{l-1}_{enc}) \tag{2}$$

$$Z^l_{enc} = LN(FFN(C^l_{enc}) + C^l_{enc}) \tag{3}$$

The self-attention sub-layer takes the output of previous layer as the input. Formally, the input for the self-attention sub-layer of the l-th layer is $Z^{l-1}_{enc} \in \mathbb{R}^{n \times d_m}$, where d_m is the dimension of output. Specially, $Z^0_{enc} = E$ and the output of encoder $Z = Z^N_{enc}$. In the process of computation, three matrices query $Q^l \in \mathbb{R}^{n \times d_m}$, key $K^l \in \mathbb{R}^{n \times d_m}$ and value $V^l \in \mathbb{R}^{n \times d_m}$ are obtained firstly by the linear projections from Z^{l-1}_{enc} with three different metrics $W^l_Q \in \mathbb{R}^{d_m \times d_m}$, $W^l_K \in \mathbb{R}^{d_m \times d_m}$ and $W^l_V \in \mathbb{R}^{d_m \times d_m}$. Then the pre-output of self-attention sub-layer can be computed with the scaled dot-product attention mechanism:

$$\begin{aligned} ATT_{self}(Z^{l-1}_{enc}) &= att(Q^l, K^l, V^l) \\ &= softmax(\frac{Q^l (K^l)^\top}{\sqrt{d_m}}) V^l \end{aligned} \tag{4}$$

Moreover, the self-attention sub-layer is normally further extended into multi-head manner. Formally,

$$\begin{aligned} ATT_{self}(Z^{l-1}_{enc}) &= Concat(H_1, \cdots, H_h) W^l_C \\ where \quad H_i &= att(Q^l (W^l_Q)_i, K^l (W^l_K)_i, V^l (W^l_V)_i) \end{aligned} \tag{5}$$

where h is the number of heads, $(W^l_Q)_i \in \mathbb{R}^{d_m \times d_h}$, $(W^l_K)_i \in \mathbb{R}^{d_m \times d_h}$, $(W^l_V)i \in \mathbb{R}^{d_m \times d_h}$ and $W^l_C \in \mathbb{R}^{h * d_h \times d_m}$ are four learnable weight matrices, d_h is the dimension for each head, we set $d_h = d_m / h$. Note that we use the multi-head attention mechanism in all the scenarios with attention method of this paper by default.

3.2.2 Emotion Sequence Decoding. The decoder in our SESGN has a similar stacked structure with N identical layers as shown in the right part of Figure 2. In addition to the two sub-layers introduced above, the decoder inserts another self-attention sub-layer in between, which performs multi-head attention over the output of the encoder. For clarity, we use the "bridge sub-layer" to refer to this additional self-attention sub-layer and $BATT(\cdot;\cdot)$ to represent the

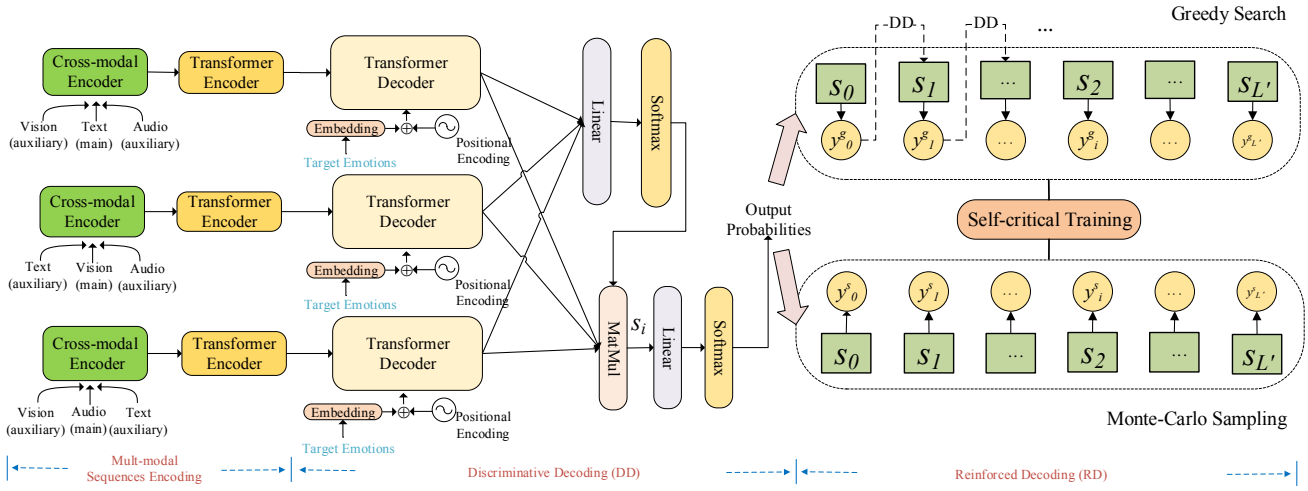

Figure 3: Multi-modal Emotion Set Generation Network (MESGN) based on Multiple Transformers.

pre-output of this sub-layer. Formally,

$$C^l_{dec} = LN(ATT_{self}(Z^{l-1}_{dec}) + Z^{l-1}_{dec}) \quad (6)$$

$$D^l_{dec} = LN(BATT(C^l_{dec}, Z^N_{enc}) + C^l_{dec}) \quad (7)$$

$$Z^l_{dec} = LN(FFN(D^l_{dec}) + D^l_{dec}) \quad (8)$$

where the calculation of $BATT(C^l_{dec}, Z^N_{enc})$ is similar to the Eq.(4) and (5), but the Q^l is obtained by linear projections from C^l_{dec} and K^l, V^l are derived from Z^N_{enc}. Especially, Z^0_{dec} is the encoded partial generated emotions and $Z^N_{dec} = S$.

Finally, for the i-th emotion decoding step, we compute a distribution over emotion category space $\mathcal{L}$ for target emotion y_i by projecting the i-th step output of decoder s_i via a linear layer with weights $W^o \in \mathbb{R}^{d_{dec} \times L}$ and a mask vector $I_i \in \mathbb{R}^L$,

$$p_i = p(y_i|y_1, \cdots, y_{i-1}; X) = softmax(W^o s_i + I_i) \quad (9)$$

where I_i is the mask vector at decoding step i that is used to prevent the decoder from predicting the repeated emotion labels. Formally,

$$(I_i)_k = \begin{cases} -\infty & \text{if label } emo_k \text{ has been predicted} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

In inferring stage, the predicted emotion label $\hat{y}_i$ for time step i is obtained by:

$$\hat{y}_i = argmax(p_i) \quad (11)$$

Moreover, $\hat{y}_i$ is further used as input token of the next generation step to predict the $(i+1)$-th emotion label. The searching repeats until the predicted emotion label is *EOS*, the mark of end-of-set.

3.3 Multi-modal Emotion Set Generation

In this section, we introduce our proposed MESGN to MMED task. The over architecture is shown in Figure 3, which consists of multi-modal encoders from three channels and multi-modal decoders from three channels. The multi-modal encoders aim to capture the cross-modal interactions among different modalities and temporal information for a better emotion generation, in which one channel of encoders includes a cross-modal transformer and a standard transformer [21]. The multi-modal decoders aim to handle both the modality-to-label and label-to-label dependencies by the discriminative decoding mechanism and reinforced decoding mechanism, respectively.

3.3.1 Multi-modal Sequences Encoding. We employ a novel cross-modal transformer to capture the cross-modal interactions between two different modalities in a many vs. many manner inspired by [19]. Then, similar to the single-modal sequence encoding, we leverage a standard transformer to capture the temporal interactions for each sequence from the above cross-modal interactions. In the following, we mainly introduce the cross-modal interaction mechanism, which is from the cross-modal attention to the cross-modal transformer, while the standard transformer encoder can refer to the above section.

Cross-modal Attention. We consider two different modalities α and β (such as facial expression and spoken words), with two (potentially unaligned) sequences denoted by $X_\alpha \in \mathbb{R}^{n_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{n_\beta \times d_\beta}$, respectively. For the rest of the paper, $n_{(.)}$ and $d_{(.)}$ are used to represent sequence length and feature dimension, respectively.

We define the Querys as $Q_\alpha = X_\alpha W_{Q_\alpha}$, Keys as $K_\beta = X_\beta W_{K_\beta}$, and Values as $V_\beta = X_\beta W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are the trainable weights. The latent interactions from β to α is presented as the cross-modal attention $Z_\alpha = CM_{\beta \to \alpha}(X_\alpha, X_\beta) \in \mathbb{R}^{n_\alpha \times d_v}$:

$$\begin{aligned} Z_\alpha &= CM_{\beta \to \alpha}(X_\alpha, X_\beta) \\ &= softmax(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}) V_\beta \\ &= softmax(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}) X_\beta W_{V_\beta} \end{aligned} \quad (12)$$

where Z_α has the same length as Q_α (i.e., n_α), but is meanwhile represented in the feature space of V_β. Specifically, the scaled (by $\sqrt{d_k}$) softmax computes a score matrix $softmax(.) \in \mathbb{R}^{n_\alpha \times n_\beta}$, whose

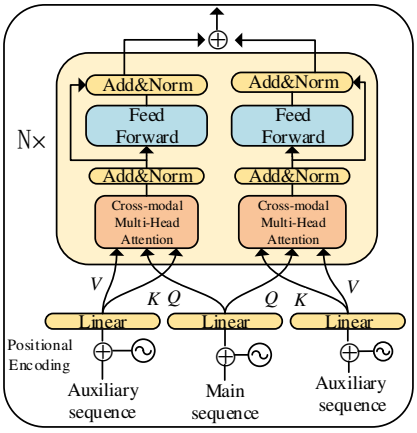

Figure 4: The architecture of cross-modal transformer encoder.

(i, j)-th entry measures the attention given by the i-th time step of modality α to the j-th time step of modality β. Hence, the i-th time step of Z_α is a weighted summary of V_β, with the weight determined by i-th row in $softmax(.)$.

Cross-modal Transformer Encoder. To make the cross-modal attention work smoothly with dot-products, we linearly transform the different inputs of three modalities into three sequences with the same dimension (not the same length for unaligned data). Besides, equipped with the positional embeddings, original inputs of three modalities are transformed to $Z^0_{\{Te,Vi,Ac\}} \in \mathbb{R}^{n_{\{Te,Vi,Ac\}} \times d_n}$.

Based on the cross-modal attention module, we employ the cross-modal transformer encoder that enables one modality for receiving information from another modality. The architecture is shown in Figure 4. In the following, we use the example for passing vision (Vi) information to text (Te), which is denoted by $Vi \rightarrow Te$. Each cross-modal transformer consists of N layers of cross-modal attention module. Formally,

$$\begin{aligned} Z^0_{Vi\rightarrow Te} &= Z^0_{Te} \\ \hat{Z}^l_{Vi\rightarrow Te} &= CM^l_{Vi\rightarrow Te}(LN(Z^{l-1}_{Vi\rightarrow Te}), LN(Z^0_{Vi})) + LN(Z^{l-1}_{Vi\rightarrow Te}) \\ Z^l_{Vi\rightarrow Te} &= FFN(LN(\hat{Z}^l_{Vi\rightarrow Te})) + LN(\hat{Z}^l_{Vi\rightarrow Te}) \end{aligned} \quad (13)$$

where $CM^l_{Vi\rightarrow Te}$ means the cross-modal attention from vision to text at layer l, which is default the multi-head version. Note that d_n should be divisible by the number of heads).

At each layer of the cross-modal attention module, the low-level signals from source modality (e.g., Vision) are transformed into a different set of Key/Value pairs to interact with the target modality (e.g., Text). Therefore, towards each modality (i.e., $Te/Vi/Ac$), we have two directional cross-modal transformers (e.g., $Vi \rightarrow Te$ and $Ac \rightarrow Te$). Subsequently, they are concatenated as a hybrid representation for a specific modality (e.g., Z^N_{Te}).

3.3.2 Emotion Set Decoding. Different from single-modal emotion set generation, we design a novel discriminative and reinforced decoding mechanism, which can simultaneously handle the dual kinds of dependencies mentioned in the introduction. On the one hand, to model the modality-to-label dependency, three transformer-based decoders from three modalities and a modality-to-label attention block are built to adaptively consider the different contributions of three modalities for each potential emotion. On the other hand, to model the label-to-label dependency, we employ a self-critic approach by reinforcement learning to generate each potential emotion conditioned on other potential emotions but not rely on the pre-defined label order.

Discriminative Decoding. Three decoders of three modalities are all based on the standard transformer decoder structure and take the previously predicted emotions and the output $Z^N_{enc,\{Te,Vi,Ac\}}$ of modality-specific encoder as the input, respectively. On this basis, at step i, each channel can obtain a final output, i.e., s^{Te}_i, s^{Vi}_i and s^{Ac}_i. We use the modality-to-label attention to weigh the different contributions from multiple modalities to the i-th potential emotion. Formally,

$$score_i = softmax([s^{Te}_i, s^{Vi}_i, s^{Ac}_i]W_s) \quad (14)$$

$$s_i = score_i \cdot [s^{Te}_i, s^{Vi}_i, s^{Ac}_i] \quad (15)$$

where $W_s \in \mathbb{R}^{d_{dec} \times 1}$ and $score_i \in \mathbb{R}^3$ denotes the dynamic contribution score from different modalities to the i-th predicting emotion. The vector s_i is considered as the effective multi-modal emotion representation to generate the i-th label. Formally,

$$p_i = p(y_i|y_1, \cdots, y_{i-1}; X^{Te}, X^{Vi}, X^{Ac}) = softmax(W^o s_i + I_i) \quad (16)$$

where I_i is the mask vector at decoding step i defined as Eq. (10).

Reinforced Decoding. Although traditional sequence decoding could handle the label-to-label dependence by conditional generation, it must rely on prior knowledge of emotion label order. This inevitably encounters the wrong penalty problem in that emotion target is a set. Therefore, we leverage the self-critic reinforcement learning by optimizing a specific metric [14, 16, 26] to measure the difference between a complete predicted label set and the real label set, rather than as the pre-defined order.

From the perspective of reinforcement learning, our emotion set reinforced decoding module can be viewed as an agent. At time-step i, the previous generated emotion labels $(y_0, \cdots, y_{i-1})$ represent the state of the agent. Defined by the parameters θ of the reinforced decoding module, a stochastic policy decides the action, determining which one is the prediction of the next emotion label. Once a complete sequence $\boldsymbol{y}$ is generated, ending with EOS, the decoding agent will observe a reward r and feedback into the model. The training objective is to minimize the negative expected reward. Formally

$$\mathcal{J}(\theta) = -\mathbb{E}_{\boldsymbol{y}} \sim p_\theta[r(\boldsymbol{y})] \quad (17)$$

According to the self-critic policy gradient algorithm, the expected gradient of the objective can be approximated as:

$$\nabla_\theta \mathcal{J}(\theta) \approx -[r(\boldsymbol{y}^s) - r(\boldsymbol{y}^g)] \nabla_\theta log(p_\theta(\boldsymbol{y}^s)) \quad (18)$$

where $\boldsymbol{y}^s$ is the sampled emotion label sequence with Monte-Carlo from the probability distribution p_θ and $\boldsymbol{y}^g$ is the generated label sequence using the greedy search algorithm. $r(\boldsymbol{y}^g)$ is our baseline, which is used to reduce the variant of gradient assessment and focus on enhancing the consistency of model training. Note that this objective function and gradient are computed over one sample.

Towards the reward, we believe that an ideal reward should be not only able to measure whether the predicted emotion labels are correct or not, but also alleviate the potential wrong penalty.

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

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	-	-

Therefore, we define the reward r as F_1 calculated by comparing generated labels with ground truth labels

$$r(\boldsymbol{y}) = F_1(\boldsymbol{y}, \boldsymbol{y}^*) \tag{19}$$

where $\boldsymbol{y}$ and $\boldsymbol{y}^*$ denote emotion label set and real emotion label set respectively. The F_1 is calculated by converting $\boldsymbol{y}$ and $\boldsymbol{y}*$ to L dimensional sparse vectors.

4 EXPERIMENTATION

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

4.1 Experimental Settings

Dataset. We use the largest available multi-modal emotion benchmark dataset, i.e., CMU-MOSEI in our evaluation. The dataset has 3,229 full-long videos which are segmented into 22856 utterance-level video clips. Each utterance-level video is annotated with multiple emotions according to three modalities, i.e., the textual, visual, and acoustic modalities. The emotion categories contain *happiness*, *sadness*, *anger*,*disgust*, *fear* and *surprise*. The average words of utterance-level video clips are 19.1. Table 1 shows the statistics of the samples with multiple labels. The average number of emotion labels per sample is 1.6. The sizes of training, validation, and test are approximately 16K, 2K, and 4.6K, respectively, as same as the split utterance-level video clips in the public SDK[1]. For the experiments with aligned data, we perform the word-level alignment following [31]. For the experiments with unaligned data, we retain the originally extracted features at word/frame level.

Implementation Details. In MESGN, we set the number of heads as 8 in multi-head crossmodal attention mechanism and multi-head attention mechanism of both the encoder and the decoder side. Following [31], 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 size of the hidden layer in the cross-modal encoder is 256, while the size in the other encoders and decoders is 512.

Besides, we make use of the dropout regularization [18] to avoid overfitting and clip the gradients [13] 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.

During the training stage, we pre-train the model for 20 epochs via MLE method. We use the Adam [10] optimization method to minimize the loss over the training data with batch size 64. For the hyper-parameters of the Adam optimizer, we set the learning rate as 0.001 for pre-training and 0.00001 for reinforcement learning with two momentum parameters of β_1 and β_2, 0.9 and 0.999 respectively. During testing, we employ the beam search algorithm [23] to find the top-ranked prediction path with the beam size 5. To motivate future development, all codes and saved models will be released via github [2].

Evaluation Metrics. In our study, we employ five 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*), micro F_1 measure (F_1), Precision (P), and Recall (R). These metrics have been popularly used in some multi-label classification problems [11, 24, 26].

Note that smaller Hamming Loss corresponds to better classification quality, while larger Accuracy, Precision, Recall, and F_1 measure corresponds to better classification quality.

4.2 Baselines

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

First, 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 simply concatenated as a new input. (1) **BR**[3] [17], which transforms the multi-label task into multiple single-label binary classification problem by ignoring the correlations between labels. (2) **CC**[3] [15], which transforms the multi-label task into a chain of binary classification problem and takes high-order label correlations into consideration. (3) **SGM**[4] [27], which considers the multi-label text classification task as a sequence generation problem with a global label embedding. (4) **LSAN**[5] [25], which takes advantage of label semantic information to determine the semantic connection between labels and documents for constructing label-specific document representation. This approach is considered as the state-of-the-art in multi-label text classification. (5) **ML-GCN**[6] [3], 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.

Second, 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 with *sigmoid* activation is used to predict the emotions. (6) **GMFN**[1] [31], which explicitly models the multi-modal interactions by capturing uni-modal, bi-modal, and tri-modal interactions. (7) **RAVEN**[7] [22], which models the fine-grained structure of nonverbal subword

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

[2] https://github.com/MANLP-suda/MMESGN

[3] http://scikit.ml/

[4] https://github.com/lancopku/SGM

[5] https://github.com/EMNLP2019LSAN/LSAN/

[6] https://github.com/Megvii-Nanjing/ML-GCN

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

Table 2: Performance of different approaches to multi-modal multi-label emotion detection with aligned and unaligned data.

Approach	Aligned					Approach	Unaligned				
	Acc	*HL*	F_1	*P*	*R*		*Acc*	*HL*	F_1	*P*	*R*
BR [17]	0.222	0.371	0.386	0.309	0.515	Average+BR	0.233	0.364	0.404	0.321	0.545
CC [15]	0.225	0.377	0.386	0.306	0.523	Average+CC	0.235	0.367	0.404	0.320	0.550
LP [20]	0.159	0.426	0.286	0.231	0.377	Average+LP	0.185	0.417	0.317	0.252	0.427
SGM [27]	0.455	0.193	0.523	0.595	0.467	CTC+SGM	0.449	0.196	0.524	0.584	0.476
LSAN [25]	0.393	0.209	0.501	0.550	0.459	CTC+LASN	0.403	0.198	0.514	0.582	0.460
ML-GCN [3]	0.411	0.207	0.509	0.546	0.476	CTC+ML-GCN	0.437	0.199	0.524	0.573	0.482
GMFN [31]	0.396	0.195	0.517	0.595	0.457	CTC+GMFN	0.386	0.212	0.494	0.534	0.456
RAVEN [22]	0.416	0.195	0.517	0.588	0.461	CTC+RAVEN	0.403	0.186	0.511	0.633	0.429
MulT [19]	0.445	0.190	0.531	**0.619**	0.465	MulT	0.423	0.184	0.523	**0.636**	0.445
MESGN (Ours)	**0.494**	**0.180**	**0.561**	0.525	**0.603**	MESGN (Ours)	**0.492**	**0.183**	**0.560**	0.530	**0.592**
MESGN w/o DD	0.487	0.214	0.552	0.535	0.570	MESGN w/o DD	0.491	0.217	0.557	0.467	0.691
MESGN w/o RD	0.470	0.194	0.544	0.582	0.511	MESGN w/o RD	0.469	0.194	0.549	0.581	0.519

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. (8) **MulT**[8] [19], 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 detection.

Third, the baselines are ablated approaches without one special component of our full model. (9) **MESGN w/o DD**, a variation of our approach, which removes discriminative decoding (multi-modal decoders) and employs only one decoder. (10) **MESGN w/o RD**, a variation of our approach, which removes the reinforced decoding (self-critic) and relies on the cross-entropy with a pre-defined label order from frequent to rare.

4.3 Experimental Results and Comparison

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 on aligned and unaligned multi-modal sequences.

For aligned data, we can observe that: 1) The classical multi-label approaches **BR**, **CC** and **LP** perform much worse than the deep learning baselines **SGM**, **LSAN** and **ML-GCN**. This indicates that the approaches with deep representation do have more advantages than the classical approaches towards multi-label problems. 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 **SGM** except *Acc*, and outperforms **LSAN** 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 **MESGN** performs best in terms of the three main metrics *Acc*, *HL*, and F_1. The t-test demonstrates that our approach significantly outperforms the best-performed baseline **MulT** ($p\text{-}value < 0.05$).

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

Table 3: Performance of single-modal (SESGN) and multi-modal (MESGN) emotion set generation approaches on the aligned data. Te: Text, Vi: Vision, Ac: Audio.

Approach	Modality	*Acc*	*HL*	F_1	*P*	*R*
SESGN	Te	0.444	0.204	0.513	**0.563**	0.471
	Vi	0.438	0.209	0.472	0.551	0.413
	Ac	0.440	0.206	0.481	0.559	0.423
MESGN	Te&Vi&Ac	**0.494**	**0.180**	**0.561**	0.525	**0.603**

For unaligned data, we utilize the average or connectionist temporal classification (CTC) [6] for the baselines to handle the multi-modal unaligned problem and perform multi-label emotion detection. We can observe that: 1) All approaches with unaligned sequences perform a little worse than those with aligned sequences, including our proposed **MESGN**. This highlights the importance of word-level alignment in multi-modal emotion detection. 2) Overall, our approach still performs best compared with all the baselines. The t-test demonstrates that our approach significantly outperforms the best-performed baseline **MulT** ($p\text{-}value < 0.05$). 3) Different from the traditional approaches of multi-modal emotion detection, our proposed **MESGN**, which models both the modality-to-label and label-to-label dependencies can well handle the multi-label problem in both multi-modal aligned and unaligned data.

Ablation study. To further demonstrate the importance of modeling modality-to-label and label-to-label dependencies, we do not model either the modality-to-label (**MESGN w/o DD**) or the label-to-label dependency (**MESGN w/o RD**) with multi-modal aligned and unaligned data, respectively. From Table 2, we observe that not modeling one of the dependencies significantly decreases the performance. This illustrates the effectiveness of our approach in modeling the two types of dependencies. Interestingly, the two types of dependencies are not equally important, giving that **MESGN w/o DD** outperforms **MESGN w/o RD** in terms of *Acc* and F_1 measure while falls behind the latter in hamming loss.

Single-modal approach vs. multi-modal approach. To illustrate the necessity of multi-modal approach for multi-label emotion detection, we also evaluate **SESGN** approach which aims at modeling a single modality while ignoring the other two. Table 3 compares

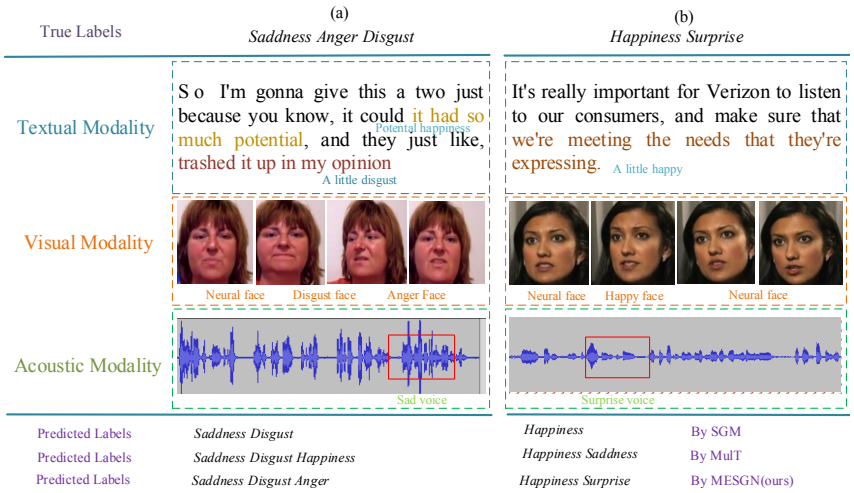

Figure 5: Two cases with labels predicted by SGM, MulT and MESGN.

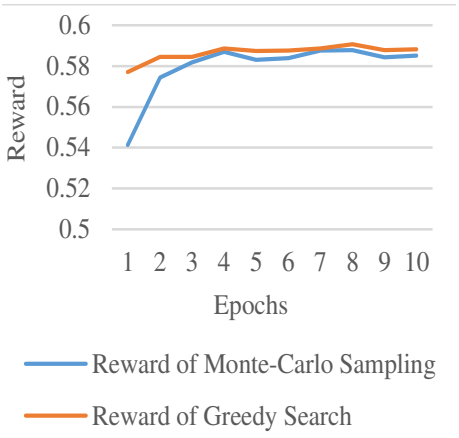

Figure 6: Reward tendency as epochs.

the performance of **SESGN** and **MESGN** approaches. From this table, we observe that **SESGN** with textual modality outperforms the counterparts with the other two modalities, suggesting that the textual modality contains more useful information for the MMED task than the others. Moreover, our **MESGN** achieves the highest performance, suggesting that both the visual and the acoustic modalities could be useful complements to the textual modality. This is consistent with our motivation that different modalities play different roles in emotion expression.

4.4 Analysis

Case study. To further demonstrate the effectiveness of our multi-modal emotion set generation network, Figure 5 presents two cases with predicted emotions by our proposed **MESGN**, and two representative baselines: one ignores modality-to-label dependency **SGM** and one ignores label-to-label dependency **MulT**. Both baselines hold the competitive performance against uni-modal multi-label approaches and multi-modal approaches: one ignores modality-to-label dependency and one ignores label-to-label dependency, respectively. From the case (a), we can see that although **SGM** can accurately detect two emotions of the ground-truth, it leaves the *Anger* emotion. This is mainly because early fusion without modality-to-label dependence results in different modalities information confusion so that it may be difficult for **SGM** to infer all the correct emotions. In contrast, **MulT** can obtain three emotions, but it only correctly detect the *Sadness* and *Disgust* emotion, and gives a wrong prediction of *Happiness*. Obviously, *Happiness* and *Sadness* are the conflicting emotions. This indicates that **MulT** is not able to capture the label-to-label dependency so that brings in the extra conflicting emotion labels.

Reward analysis. Figure 6 shows the reward tendency of Monte-Carlo sampling (MC) and greedy search (GS). From this figure, we can see that the reward generated by MC performs lower than that by GS before the third epoch, indicating MC in the exploration state at previous stage. Then, it is gradually equal to GS. This shows that the reinforced decoding mechanism can learn stably and converge quickly.

5 CONCLUSION

Conventional approaches for multi-modal multi-label emotion detection (MMED) normally perform the binary classification within each emotion category. This obviously neglects both the modality-to-label and label-to-label dependencies. In order to model both kinds of dependencies simultaneously for MMED, we propose a multi-modal emotion set generation network (MESGN). Based on the multi-modal transformer encoding sub-network, MESGN contains a novel decoding sub-network, in which the discriminative decoding module can effectively model the modality-to-label dependency, and the reinforced decoding module can effectively model the label-to-label dependency. Specifically, we first construct three cross-modal transformer encoders and obtain an interactive sequence representation for each modality. Then, the interactive sequence goes through a standard transformer encoder to capture temporal information. On this basis, we feed the outputs of the three transformer encoders into three discriminative transformer decoders and introduce the modality-to-label attention to control the contribution of different modalities to each potential emotional label. Meanwhile, we introduce a reinforced decoding mechanism to free the dependence of label order in the traditional sequence generation but predicting a potential emotion conditioned on other potential emotions. Detailed evaluations on both the aligned and unaligned multi-modal data demonstrate the effectiveness of our proposed MESGN against the state-of-the-art in multi-modal emotion detection and multi-label classification.

6 ACKNOWLEDGMENTS

The research work is partially supported by the Key Project of NSFC No. 61751206 and two NSFC grants No. 61673290, No.61876120. This work is also supported by a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

REFERENCES

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

[2] 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-IJCNLP 2019*. Association for Computational Linguistics, Hong Kong, China.

[3] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition With Graph Convolutional Networks. In *Proceedings of IEEE-CVPR 2019*. 5177–5186.

[4] Maros Galik and Stefan Rank. 2012. Modelling Emotional Trajectories of Individuals in an Online Chat. In *Proceedings of MATES 2012*. 96–105.

[5] 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*. 154–164.

[6] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML 2006*. 369–376.

[7] 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*. 157–166.

[8] Neel Kant, Raul Puri, Nikolai Yakovenko, and Bryan Catanzaro. 2018. Practical Text Classification With Large Pre-Trained Language Models. *arXiv preprint arXiv:1812.01207* (2018).

[9] 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*. 141–145.

[10] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR 2015*.

[11] Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level Emotion Classification with Label and Context Dependence. In *Proceedings of ACL 2015*. 1045–1053.

[12] 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*. 6818–6825.

[13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of ICML 2013*. 1310–1318.

[14] Marc'Aurelio Ranzato, Sumit Chopra andMichael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *Proceedings of ICLR 2016*.

[15] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Mach. Learn.* 85, 3 (2011), 333–359.

[16] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning. In *Proceedings of IEEE-CVPR 2017*. 1179–1195.

[17] 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*. 188–199.

[18] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958.

[19] 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*. Association for Computational Linguistics, 6558–6569. https://doi.org/10.18653/v1/p19-1656

[20] Grigorios Tsoumakas and Ioannis Katakis. 2009. Multi-Label Classification. In *Proceedings of IGI 2009*. IGI Global, 309–319.

[21] 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*. 6000–6010.

[22] 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*. 7216–7223.

[23] Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-Sequence Learning as Beam-Search Optimization. In *Proceedings of EMNLP 2016*. 1296–1306.

[24] 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-IJCNLP 2019*. Association for Computational Linguistics, 4353–4363.

[25] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-Specific Document Representation for Multi-Label Text Classification. In *Proceedings of EMNLP-IJCNLP 2019*. Association for Computational Linguistics, 466–475.

[26] 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*. 5252–5258.

[27] 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*. 3915–3926.

[28] 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@EMNLP 2019*. 316–321.

[29] 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*. 1097–1102.

[30] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Proceedings of AAAI 2018*. 5634–5641.

[31] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of ACL 2018*. 2236–2246.

[32] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In *Proceedings of AAAI 2018*. 5642–5649.

[33] Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Effective Sentiment-relevant Word Selection for Multi-modal Sentiment Analysis in Spoken Language. In *Proceedings of ACM MM 2019*. ACM, 148–156.

[34] Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling the Clause-Level Structure to Multimodal Sentiment Analysis via Reinforcement Learning. In *Proceedings of IEEE ICME 2019*. IEEE, 730–735.

[35] Dong Zhang, Liangqing Wu, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Multi-Modal Language Analysis with Hierarchical Interaction-Level and Selection-Level Attentions. In *Proceedings of IEEE ICME 2019*. IEEE, 724–729.

[36] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In *Proceedings of IJCAI 2019*. ijcai.org, 5415–5421.

[37] Xiabing Zhou, Zhongqing Wang, Shoushan Li, Guodong Zhou, and Min Zhang. 2019. Emotion Detection with Neural Personal Discrimination. In *Proceedings of EMNLP 2019*. 5502–5510.