Unimodal and Crossmodal Refinement Network for Multimodal Sequence Fusion

Xiaobao Guo[1,2], Adams Kong[1], Huan Zhou[3], Xianfeng Wang[3], and Min Wang[3]
[1] School of Computer Science and Engineering, Nanyang Technological University
[2] Rapid-Rich Object Search (ROSE) Lab, IGP, Nanyang Technological University
[3] AI Application Research Center (AARC), Huawei Technologies
xiaobao001@e.ntu.edu.sg, adamskong@ntu.edu.sg
{zhou.huan,wangxianfeng10,wangmin5}@huawei.com

Abstract

Effective unimodal representation and complementary crossmodal representation fusion are both important in multimodal representation learning. Prior works often modulate one modal feature to another straightforwardly and thus, underutilizing both unimodal and crossmodal representation refinements, which incurs a bottleneck of performance improvement. In this paper, Unimodal and Crossmodal Refinement Network (UCRN) is proposed to enhance both unimodal and crossmodal representations. Specifically, to improve unimodal representations, a unimodal refinement module is designed to refine modality-specific learning via iteratively updating the distribution with transformer-based attention layers. Self-quality improvement layers are followed to generate the desired weighted representations progressively. Subsequently, those unimodal representations are projected into a common latent space, regularized by a multimodal Jensen-Shannon divergence loss for better crossmodal refinement. Lastly, a crossmodal refinement module is employed to integrate all information. By hierarchical explorations on unimodal, bimodal, and trimodal interactions, UCRN is highly robust against missing modality and noisy data. Experimental results on MOSI and MOSEI datasets illustrated that the proposed UCRN outperforms recent state-of-the-art techniques and its robustness is highly preferred in real multimodal sequence fusion scenarios. Codes will be shared publicly[1].

1 Introduction

Motivated by recent research achievements on modality representations in language (Pennington et al., 2014; Devlin et al., 2019; Brown et al., 2020), audio (Degottex et al., 2014; Chen et al., 2018; Li et al., 2019), and vision (He et al., 2016; Woo et al., 2018; Li and Deng, 2020), multimodal learning (Baltrušaitis et al., 2018) that aims to efficiently extract joint representations from multiple sensory data, has drawn much attention recently. By mining complementary relations across modalities, multimodal fusion could present a more reliable and comprehensive interpretation of the world.

With the development of various network architectures, considerable progress has been achieved in multimodal fusion (Williams et al., 2018; Tsai et al., 2019; Mai et al., 2020; Zadeh et al., 2019), showing that multimodal representation outperforms unimodal ones on emotion and sentiment prediction tasks. Additionally, recent works seek to improve the efficacy of the multimodal fusion methods by assuming that one modality can be translated to another (Tsai et al., 2019; Mai et al., 2020), or modulated by pivot modality (Delbrouck et al., 2020), so as to align pairwise representations in a common space.

However, converting one modality to another modality appears to be inadequate in projecting all modalities into one feature space. A specific pattern may not coexist in all modalities (*e.g.*, it expresses happiness in language/audio while showing neutrality on facial expression). Also, the conversion is usually followed by a downstream fusion to produce the final fusion result. Therefore, the downstream fusion may have already encompassed it as a byproduct making conversion redundant in terms of multimodal fusion tasks.

It has been frequently reported that lexical representation is a stronger predictor than audio and vision representations. Hence, (Delbrouck et al., 2020) leverages language feature to modulate others. As a result, it becomes the main contributor towards multimodal fused representation. On the contrary, through this process, those weak predictors are prone to undermine the modality-common representations. From this view, the strategy of direct modality modulation or pairwise translation

[1] https://github.com/HW-AARC-CLUB/emnlp2021_ucrn

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9143–9153
November 7–11, 2021.

may jeopardize learning rich-fused representation and lead to suboptimal results. Moreover, most of those network architectures require all modalities as input. As a result, the learned representations may perform poorly in the real world where complete modalities might not always be simultaneously available (*e.g.*, some specific modality is missing or noisy). This may be because of over fusing or losing sight of addressing the importance of unimodal refinement.

Although the presence of multiple modalities provides additional information, there are two key challenges to be addressed when learning from multimodal data: 1) models must learn the complex intramodal and crossmodal interactions for predictions, and 2) models must be robust to unexpected missing or noisy modalities during testing.

To address the aforementioned problems, the Unimodal and Crossmodal Refinement Network (UCRN) is proposed, which takes both robust unimodal representation and efficient crossmodal representation into consideration. Our hypotheses are 1) a robust unimodal representation is essential for efficient multimodal fusion, 2) it is beneficial to reduce modality gaps before modality fusion, and 3) stacks of attention-based mechanism can efficiently select the most salient features within a representation, favoring robust representation learning.

Guided by the above hypotheses, the proposed UCRN consists of the following sub-modules with inputs of basic modality sequence representations: 1) a unimodal refinement module is proposed to yield robust modality-specific representations; 2) the robust unimodal representations are projected to different latent spaces with the aims of reducing representation gaps and 3) the latent spaces are concatenated to correlate modality-common features and to produce robust multimodal representation. The final predicted sequence label is the one with the highest probability in the network output.

To recap, the contributions of this paper can be summarized as follows:

- Unimodal and Crossmodal Refinement Network (UCRN) is proposed to perform robust and efficient multimodal representation learning;
- The unimodal representation refinement is proposed for improving crossmodal fusion;
- The crossmodal refinement is explored to reduce modality gaps by progressively refining and fusing the flexibly concatenated unimodal representations;
- Experiments are conducted on widely studied multimodal datasets. The results demonstrate that, compared with recent state-of-the-art (SOTA) works, the UCRN shows competitive performances and strong robustness against noisy and absent some unimodal inputs.

2 Related Work

2.1 Multimodal Sequence Learning

Multimodal sequence learning extracts the inter- and isntra-dependencies of multimodal data and uses complementary information to improve model performance. Many methods captured the sequential information by taking the advantage of LSTM (Hochreiter and Schmidhuber, 1997). (Zadeh et al., 2018a) used a recurrent model for multimodal learning, where a dynamic memory module was proposed for crossmodal interaction learning. MARN (Zadeh et al., 2018b) learned multimodal information by a hybrid LSTM component while MV-LSTM (Rajagopalan et al., 2016) computed representations for each modality inside a multimodal LSTM variant. Mu-Net (Shenoy and Sardana, 2020) took context information into account and adopted a recurrent model to learn the dependency among speakers. However, LSTM is difficult to train, and certain intermediate information may cause interference or bring conflicts in fusion, thus leading to unsatisfactory results.

To extract complementary information and to perform multimodal sequence fusion, many methods have investigated the fusion strategies. Early fusion performed fusion at the input level by simply concatenating multimodal features (Morency et al., 2011; Pérez-Rosas et al., 2013; Poria et al., 2016), which did not model the intra-modality dynamics efficiently because unimodal representations can be complicated and are not easy to learn in the whole model, as well as posing the potential of overfitting. Late fusion approaches trained unimodal classifiers individually and made fusion by voting (Wang et al., 2016; zad). Although this made intra-dynamic modeling more effective, simply applying a weighted average might not produce the best fusion results.

2.2 Translation-based Method

To better model the interactions among multimodal sequences, translation-based methods (Tsai et al.,

2019; Pham et al., 2019; Mai et al., 2020) assumed that the representation of one modality can be converted to another, thus minimizing the gap between unimodal representations. For example, MulT (Tsai et al., 2019) proposed a multimodal transformer architecture that translated any two modalities to the remaining one, then combined the translated features for final fusion; MCTN (Pham et al., 2019) leveraged an encoder-decoder structure to convert one modality to another, as well as using a cyclic consistency loss to produce better modality translation results. AGFN (Mai et al., 2020) was proposed to learn a common embedding space via translating a modality to a target one, which takes adversarial training and graph-based fusion mechanism for prediction. CIA (Chauhan et al., 2019) implemented translation-based fusion on contextual attention modeling, where crossmodal auto-encoding was utilized to extract features. Translation-based methods directly converted one modality representation to another, which relied upon much reference information from pairwise modalities, imposing a limit on solving the modality missing issue.

2.3 Transformer and Self-attention

Transformer (Vaswani et al., 2017) is an effective and strong network to conduct sequence modeling. Different from recurrence modeling, it shows superiority in training and performance on many tasks (Yu et al., 2019; Hu et al., 2020; Naseem et al., 2020) based on attention mechanism. It transforms one sequence to another with an encoder-decoder structure, where the attention mechanism weighs the input sequence to decide which part is important at each step. By using Transformer encoding, TBJE (Delbrouck et al., 2020) proposed monomodal and multimodal variants; yet, they were not in a unified architecture, leading to a problem of model selection. Besides, the performance degraded when modulating the added visual information by language features, which was unsatisfactory in terms of multimodal fusion.

Different from the attention mechanism in Transformer, self-attention (Hu et al., 2018) was also a technique that has been widely used to extract contextual and correlated information within features. The attention-based mechanism shows promising results of modeling sequences. Thus, it is also adopted in the proposed UCRN. However, instead of using modal translation or separated models, this paper highlights the unimodal representations refinement and the crossmodal representation refinement by regularizing the multi-modality inputs into a common space.

3 Unimodal and Crossmodal Refinement Network (UCRN)

In this paper, Unimodal and Crossmodal Refinement Network (UCRN) is proposed. As shown in Figure 1, UCRN is comprised of three main parts, where the first part conducts unimodal representation refinement, the second part refines all the previous information for fusion and learns a modality-common representation, and the last part performs prediction. The Unimodal Refinement Module (URM) takes unimodal features (*i.e.*, language, audio, vision features) as input to learn the refined unimodal representations. Then the refined unimodal representations are mapped to a common latent space by imposing a Multimodal Jenson-Shannon (MJS) divergence regularizer. Following this, the Self-Quality Improvement Layers (SQIL) are used to further extract desired weighted unimodal representations for fusion. Lastly, the Crossmodal Refinement Module (CRM) integrates all information and extracts multimodal interactions.

3.1 Problem Definition

Suppose we have the i-th input feature $\mathbb{X}^i$, $\mathbb{X}^i = \{x_m^i \in \mathbb{R}^{d_m \times t_m}; m \in \{l, a, v\}\}$, where l, a, v represents language, audio, and vision, respectively, and d_m and t_m denote the dimensions of the modal feature and time sequence, respectively. Let K be the batch size. The goal of multimodal sequence fusion is to determine a deep fusion network $F(\mathbb{X}^i)$ so that the output $\hat{y}^i$ is expected to approximate the target y^i. This can be achieved by minimizing the loss as

$$\min_F \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}(\hat{y}^i = F(\mathbb{X}^i), y^i). \quad (1)$$

3.2 Unimodal Refinement

Unimodal Refinement Module (URM) is designed to reinforce the modality-specific learning. High quality unimodal representations would benefit multimodal fusion. URM based on a transformer architecture takes a single modal feature as input to learn a robust and refined unimodal representation.

As shown in Fig. 1, U_m is corresponding to the URM for modality m. The U_m is trained to

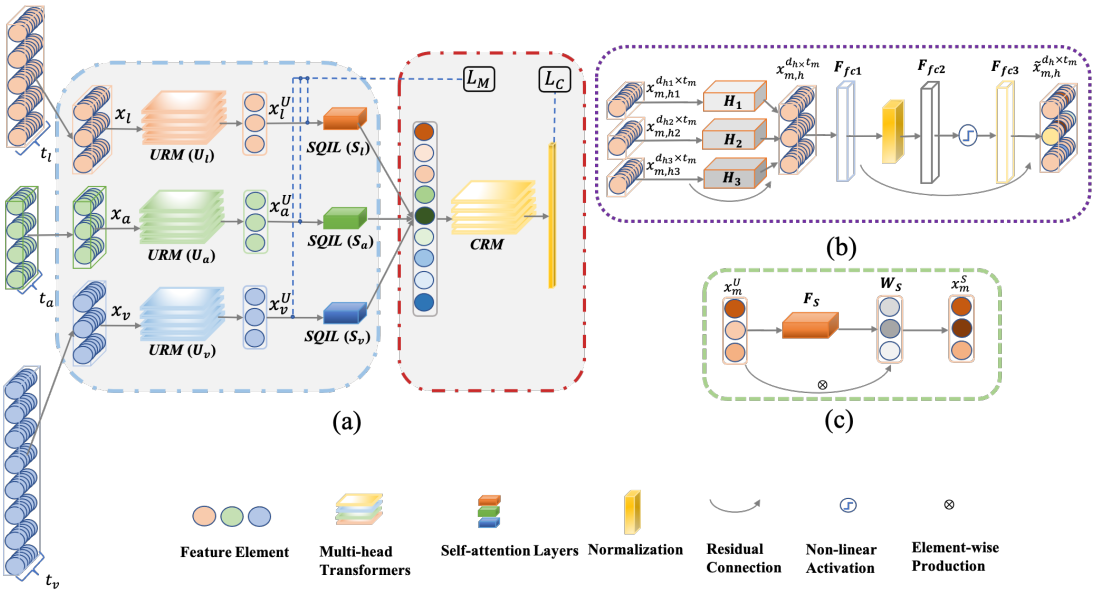

Figure 1: (a). Architecture of Unimodal and Crossmodal Refinement Network (UCRN). x_l, x_a, and x_v are input unimodal features. Unimodal Refinement Module (URM) takes unimodal features to learn the refined unimodal representations x_m^U, $m \in \{l, a, v\}$. SQIL is followed to extract desired weighted unimodal representations for fusion. Crossmodal Refinements Module (CRM) integrates all information and refines multimodal interactions for prediction. L_M and L_C are multimodal JS divergence loss and classification loss; (b). Multi-head Transformer Encoder. Unimodal representations are split to each head and then concatenated to produce activated refined results; (c). Self-quality Improvement Layers (SQIL). The results are learned weighted unimodal representations.

learn the unimodal representation with a multi-layer transformer-based network. Therefore, we have,

$$x_m^U = U_m(x_m; \theta_m), \tag{2}$$

where x_m^U represents the refined unimodal representation for modality m, while θ_m stands for the parameters of U_m. In fact, before inputting each unimodal feature to the URM, it is first sent to a projection layer to convert each feature to a specific dimension, which can simplify the subsequent unified operations. Then, the unimodal representation is passed to a multi-layer multi-head transformer. Five layers and three heads are used. Let H_k, where $k = \{1, 2, 3\}$, denote the head in each layer and $x_m^{d_m \times t_m}$ be the projected feature. As presented in Fig. 1 (b), for $x_m^{d_m \times t_m}$, d_m is evenly split to d_{hk}, which is the feature dimension in head k. The operations of the mutli-head transformer are described by the following equations,

$$x_{m,h}^{d_h \times t_m} = \oplus\{H_k(x_{m,hk}^{d_{hk} \times t_m}) + x_{m,hk}^{d_{hk} \times t_m}\}, \tag{3}$$

$$\hat{x}_{m,h}^{d_h \times t_m} = F_{fc_1}(x_{m,h}^{d_h \times t_m}), \tag{4}$$

$$\tilde{x}_{m,h}^{d_h \times t_m} = F_{fc_3}(ReLU(F_{fc_2}(LN(\hat{x}_{m,h}^{d_h \times t_m})))) + \hat{x}_{m,h}^{d_h \times t_m}, \tag{5}$$

where $F_{fc_i}(\cdot)$, $i \in \{1, 2, 3\}$, is a fully connected layer, and $LN(\cdot)$ denotes a layer normalization. $x_{m,hk}^{d_{hk} \times t_m}$ is the input and $H_k(x_{m,hk}^{d_{hk} \times t_m})$ is the output from head k. $\oplus$ is the operation of concatenation over the output from each head. URM can learn the sequence representation by extracting the last time step in the time dimension as the key step, thus the time space collapses to one. Here, we have the overall definition:

$$x_m^U = \tilde{x}_m^{d_m \times 1} = U_m(x_m; \theta_m). \tag{6}$$

3.3 Multimodal Jenson-Shannon Divergence Regularizer

Due to heterogeneity across divergent modalities, the fused multimodal representation follows an unknown yet complex distribution. In order to further enhance crossmodal refinement, we propose to regularize the distribution by explicitly adding a regularizer. It is well known that the Kullback-Leibler Divergence, $\mathcal{D}_{KL}$, can measure distribution differences. However, since the commutative consistency of a pair of modalities should be kept in our framework, the multimodal Jensen-Shannon

divergence $\mathcal{D}_M$ is employed, which is defined as:

$$\mathcal{D}_M(\alpha,\beta) = \frac{1}{2}(\mathcal{D}_{KL}(p(\alpha), \frac{p(\alpha)+p(\beta)}{2}) + \mathcal{D}_{KL}(p(\beta), \frac{p(\alpha)+p(\beta)}{2})), \tag{7}$$

where $(\alpha,\beta) \in [(l,a),(a,v),(v,l)]$, $p(\alpha)$ and $p(\beta)$ represent the probability distributions of the learned features for n classes: $p(\alpha) = \{p_1(\alpha), p_2(\alpha), \ldots, p_n(\alpha)\}$ and $p(\beta) = \{p_1(\beta), p_2(\beta), \ldots, p_n(\beta)\}$. $\mathcal{D}_M$ serves as a regularizer on x_m^U and aims to optimize the whole framework. To learn a common representation for fusion, the objective function $\mathcal{L}_M$ regularizing the probability distributions of all modalities is defined as

$$\mathcal{L}_M = \mathcal{D}_M(l,a) + \mathcal{D}_M(a,v) + \mathcal{D}_M(v,l). \tag{8}$$

3.4 Self-Quality Improvement Layers

Self-Quality Improvement Layers (SQIL) are added to further produce the desired unimodal representations for fusion. SQIL is a stack of simple self-attention layers that learns the weighted unimodal representations for fusion.

$$\begin{aligned} W_S =& F_S(x_m^U), \\ x_m^S =& S_m(x_m^U;\theta_s) = x_m^U \cdot W_S, \end{aligned} \tag{9}$$

where F_S represents the linear transformation and nonlinear activations (ReLU), and W_S is the self-attention weight matrix. x_m^U is the output obtained from the URM as shown in Fig. 1 (a).

3.5 Crossmodal Refinement

Crossmodal Refinement Module (CRM) aims to learn effective crossmodal representations by integrating all refined unimodal representations. CRM (in Fig. 1 (a)) is also built based on a multi-head transformer, which takes the concatenation of the weighted unimodal representations as input. CRM becomes a bimodal or unimodal fusion module if one or two modalities are missing. CRM adaptively captures the dynamics of multimodal interactions and extracts key information among the inputs. It is a light yet effective fusion module. Specially, to ensure the proposed modal robust to noise and missing information in any modalities, flexible modality combinations are supported herein. Let $C(x_j;\theta_c)$ be the transformation of CRM with θ_c being the parameters.

$$x_j = \oplus\, x_m^S, \tag{10}$$

where $j \in \{l,a,v,(l,a),(a,v),(v,a),(l,a,v)\}$ denotes any possible combinations from the three multimodal inputs. $\oplus$ denotes the concatenation operation and x_m^S are features obtained from Eq. (9). Note that x_m^S are from modalities in j. Then the fused representation can be represented as:

$$x_j^C = C(x_j;\theta_c). \tag{11}$$

Lastly, the fused feature x_j^C is passed to two fully connected layers $F_{fc}(\cdot)$ before performing classification or regression with a loss function of $\mathcal{L}_C$. Here, either the cross-entropy loss $\mathcal{L}_{ce}$ or the least absolute deviations $\mathcal{L}_1$ is applied for diffident learning tasks. Specifically,

$$\begin{aligned} \mathcal{L}_C =& \mathcal{L}_{ce}(F_{fc}(x_{j,c}^C), y_{j,c}) = -\sum_{c=1}^{N} y_{j,c}\log(F_{fc}(x_{j,c}^C)), \\ or \quad \mathcal{L}_C =& \mathcal{L}_1(F_{fc}(x_{j,c}^C), z_{j,c}) = \sum_{c=1}^{M} |F_{fc}(x_{j,c}^C) - z_{j,c}|, \end{aligned} \tag{12}$$

where N is the number of classes, and $y_{j,c}$ is the ground-truth label for $\mathcal{L}_{ce}$, and M is the number data and $z_{j,c}$ is the target regression value for $\mathcal{L}_1$. The final UCRN objective function is

$$\mathcal{L} = \lambda_m \mathcal{L}_M + \lambda_c \mathcal{L}_C, \tag{13}$$

where URM, SQIL, and CRM are jointly optimized, and λ_m and λ_c are two trade-off parameters for the two loss terms. They are set to 1 as default in our experiments. Extensive experiments in the following section demonstrate that UCRN not only improves the performance over multiple multimodal datasets but also is robust against modality missing and noise.

4 Experiments

4.1 Datasets

CMU-MOSI (Zadeh et al., 2016) dataset is a multimodal opinion sentiment intensity analysis dataset, which consists of 2,199 short monologue video clips (opinion utterances). There are 35 facial action units that record facial muscle movement (Ekman et al., 1980; Ekman, 1992). Low-level acoustic features are extracted by COVAREP (Degottex et al., 2014). Language data are segmented by word and expressed as discrete word embedding (Pennington et al., 2014).

CMU-MOSEI (Zadeh et al., 2018c) is a sentiment and emotion analysis dataset made up of 23,454 movie review video clips. The feature extraction methods are the same as MOSI.

Method	Sentiment			Emotion												WA
	2-class	7-class	Emot	Happy		Sad		Angry		Fear		Disgust		Surprise		
	A	A	avg A	A	F1	A	F1	A	F1	A	F1	A	F1	A	F1	
GRAPH-mfn†	76.9	**45.0**	62.35	66.3	66.3	60.4	66.9	62.6	72.8	62.0	**89.9**	69.1	76.6	53.7	**85.5**	✓
Mu-net †	82.1	-	82.77	70.0	**68.4**	76.1	**74.5**	83.1	**80.9**	89.7	87	**90.3**	**87.3**	87.4	84	✓
UCRN †	**82.25**	42.76	**83.25**	**70.8**	65.2	**77.2**	67.9	**84.4**	75.6	**90.5**	86.4	88.7	84.5	**87.9**	**85.5**	✓
TBJE*	81.5	44.4	80.68	65	64	72	67.9	81.6	74.7	89.1	84	85.9	83.6	90.5	86.1	✗
UCRN*	**84.36**	**46.84**	**82.27**	**67.3**	**67.1**	**74.5**	**68.6**	**82.8**	**76.8**	**89.7**	**88.6**	**87.8**	**85.2**	**91.5**	**86.7**	✗

Table 1: Sentiment and emotion accuracy (%) and F1 score (%) comparisons on CMU-MOSEI (marked by †) and TBJE-MOSEI (marked by *). 'A' stands for accuracy and 'F1' stands for F1 score. WA indicates weighted accuracy on the emotion task. For a fair comparison, WAs are compared with GRAPH-mfn and Mu-Net on CMU-MOSEI and standard accuracies are compared with TBJE on TBJE-MOSEI.

Method	avg F1	avg Acc	MAE	Corr
LMF	75.7	76.4	0.91	0.67
BC-LSTM	78.1	77.9	-	-
TFN	78.3	78.3	-	-
LMFN-e2e	79.4	79.4	-	-
MulT	81.0	81.1	**0.87**	**0.69**
ARGF	81.4	81.3	-	-
Mu-net	80.10	81.19	-	-
UCRN	**82.01**	**81.71**	0.89	**0.69**

Table 2: 2-class sentiment performance comparison on CMU-MOSI. Avg F1 and avg Acc are percentages (%).

Method	avg F1	avg Acc	MAE	Corr
LMF	55.80	59.41	-	-
LMFN	59.48	61.31	-	-
ARGF	59.03	60.77	-	-
GRAFH-mfn	77.0	76.9	0.71	0.54
CIA	78.23	80.37	0.68	0.55
MulT	81.6	81.6	**0.59**	**0.69**
Mu-net	80.01	82.10	**0.59**	0.50
UCRN	**82.30**	**82.25**	0.60	0.68

Table 3: 2-class sentiment performance comparison on CMU-MOSEI. Avg F1 and avg Acc are percentages (%).

TBJE-MOSEI (Delbrouck et al., 2020) is pre-processed from original CMU-MOSEI dataset but using different feature extraction methods. For fair comparision with TBJE, the experimental results are also reported on this dataset.

Metrics 2-class sentiment accuracy[2], F1 score, MAE (mean square error, the lower the better) and Corr (Correlation) are used as performance indexes. 7-class sentiment accuracy and emotion classification results are also reported for comparing several strong benchmarks.

4.2 Implementation Details

Experimental Settings All the multi-head transformer-based architectures in the URM and CRM are implemented with 5 layers and 3 heads. In each transformer encoding layer in the URM, refined unimodal representation is trained by parsing query, key, and value the same input. In SQIL, the refined unimodal representation is first passed through a global average pooling on feature dimension, and then a fully connected layer to learn the correlation within features, resulting in a weighted unimodal representation. In CRM, the query, key and value of the transform inputs are the concatenated refined multimodal representation.

Training UCRN is trained in an end-to-end manner. It is light and easy to train. The batch size is set to 16 (32) and a basic learning rate is 1e-3 (2e-3) on MOSI (MOSEI).

Test The proposed model is tested on MOSI and MOSEI for sentiments and emotions. More details are reported in the supplementary materials for readers to reproduce.

4.3 Quantitative Results Compared with Benchmarks

Extensive experiments are conducted on several multimodal sentiment analysis and emotion prediction datasets. The methods compared in this work are the state-of-the-arts, among which MulT (Tsai et al., 2019), Mu-Net (Shenoy and Sardana, 2020), and TBJE (Delbrouck et al., 2020) are strong benchmarks. MulT (Tsai et al., 2019) used explicit source-target modality translation, Mu-Net (Shenoy and Sardana, 2020) adopted a pairwise attention mechanism for fusion, and TBJE (Delbrouck et al., 2020) implicitly used one modality (language) to modulate others.

Table 1 lists both sentiment and emotion performance comparisons. Specifically, for sentiment analysis, 2-class and 7-class standard accuracies are both reported on CMU-MOSEI (marked by †) and TBJE-MOSEI (marked by *). UCRN shows competitive results on these two tasks. On the emo-

[2]Please refer to the officially released evaluation metrics from https://github.com/A2Zadeh/CMU-MultimodalSDK

Method	Acc-2 (%) (L+A/L+V/A+V)	Acc-7(%) (L+A/L+V/A+V)	Emot avg A(%) (L+A/L+V/A+V)	WA
MulT†	1.30 / 4.33 / 10.16	-	-	-
Mu-Net†	2.34 / 2.48 / 8.45	-	10.22 / 8.41 / 12.71	✓
TBJE†	-0.68 / 2.94 / 7.52	-0.49 / 5.43 / 7.10	-	-
UCRN†	**1.17 / 2.18 / 6.16**	2.69 / **4.44 / 6.10**	**1.39 / 0.79 / 2.62**	✓
TBJE*	-1.10 / 1.69 / 2.27	-2.48 / 3.27 / 6.85	-0.83 / 1.08 / 2.18	✗
UCRN*	1.19 / **0.59 / 1.53**	2.49 / **2.16 / 1.52**	0.44 / **0.29 / 1.22**	✗

Table 4: Performance percentage drop comparison by masking a certain modality on CMU-MOSEI (marked by†) and TBJE-MOSEI (marked by *). WA indicates average weighted accuracy of emotion. 'L', 'A', and 'V' are abbreviations for language, audio, and visual modalities, respectively.

Method	Acc-2(%)	Acc-7(%)	Emot avg A(%)	WA
MulT†	5.48	-	-	-
Mu-Net†	3.59	-	10.44	✓
TBJE†	2.63	5.38	-	-
UCRN†	**2.37**	**4.05**	**2.87**	✓
TBJE*	2.87	10.86	2.69	✗
UCRN*	**1.87**	**4.53**	**2.54**	✗

Table 5: Performance percentage drop comparison by adding random noise on CMU-MOSEI (marked by †) and TBJE-MOSEI (marked by *). WA indicates average weighted accuracy of emotion.

tion classification task, for a fair comparison, the weighted accuracies (WA) are reported as Mu-Net and GRAPH-mfn adopted it. The unweighted results are compared with TBJE. UCRN shows the best average accuracy for emotion classification over all compared methods.

Further comparisons on 2-class sentiment analysis are presented in Table 2 and Table 3. The compared methods include LSTM-based, translation-based, and pairwise-learning based. On both CMU-MOSI and CMU-MOSEI datasets, UCRN shows improvement over the compared methods and outperforms the strong benchmark methods in terms of average F1 score and accuracy.

The results indicate that the refined unimodal and crossmodal representations are of vital importance for multimodal fusion. UCRN shows a competitive performance owing to the ample exploration of unimodal dynamics from URM and the effective crossmodal representation from CRM. UCRN shows advantages to adaptively take any combination of input modalities. More results are presented in the following subsection.

4.4 Robustness Experiments

Missing modality and noise are ever-present in the real world. Due to non-alignment, missing, or incomplete modalities, the information expressed by unimodal features is disproportionate. Therefore, translation-based methods are inclined to become invalid in those cases. However, UCRN can alleviate these problems. Defining robustness as the percentage decrease in accuracy, *i.e.*, (trimodal accuracy - masked or noisy modality accuracy) / trimodal accuracy, it allows us to objectively evaluate the robustness of UCRN. Two kinds of experiments were conducted to validate the robustness of UCRN against modality missing and noise with several strong translation-based and pairwise mapping based benchmarks.

Firstly, to simulate missing modality, features of a modality will be masked. Under such circumstances, UCRN still achieves a higher accuracy than its counterparts, which demonstrates its robustness. As shown in Table 4, masking one of the vision, audio, or language modality, UCRN gets more robust results on the average performance. Assuming that more modalities have a greater representation capability, the case of performance degradation given added modality should not be taken into account in the robustness aspect. Note that TBJE shows a degradation with trimodal input compared to its bimodal one (*i.e.*, 2-class accuracy L+A+V of 81.5% and L+A of 82.4% according to (Delbrouck et al., 2020)). This is because TBJE cannot deal with the disparity among modalities or fully explore vision features, which results in the overall representation being impaired.

In spite of that, UCRN still outperforms TBJE in terms of accuracy (*i.e.*, 2-class accuracy L+A+V of 84.36% and L+A of 83.35%). Therefore, we only compare the performance drops especially on the cases of L+V and A+V.

Secondly, to simulate the presence of noise during information acquisition in the real world, noise that follows a Bernoulli distribution is randomly added on the entire modality features with a probability of noise presence 0.5. Results in Table 5

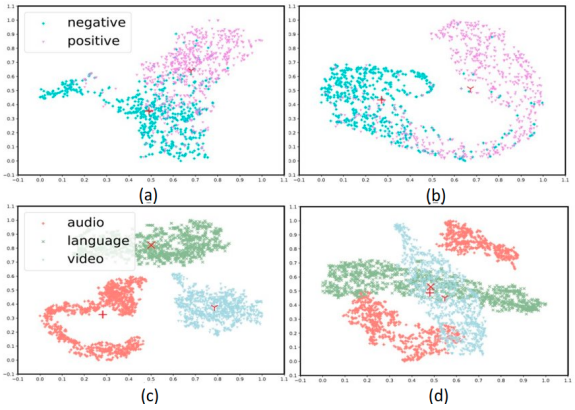

Figure 2: Visualization for distributions of multimodal features in embedding space. Please zoom in for a better view.

show that UCRN is more robust for both sentiment prediction and emotion classification tasks.

UCRN is robust against missing modality and noise because it explores the refined unimodal representation and correlates the crossmodal features adaptively.

We argue that the translation-based and implicit modulation methods have a limitation on robustness due to pairwise interactions and source-target translation.

4.5 Ablation Study

UCRN is powerful to reduce modality gap. To show this, the visualization for distributions of multimodal features in embedding space is provided in Fig. 2. The t-SNE algorithm was utilized to transform feature vectors to the 2D maps. Fig. 2 (a) shows the feature embeddings right before 2-class sentiment prediction without SQIL and without $\mathcal{L}_M$, whereas Fig. 2 (b) is obtained from UCRN with those modules. The feature embeddings in Fig. 2 (b) become more clustered and separable. Fig. 2 (c) and Fig. 2 (d) show the distributions of unimodal features before crossmodal fusion with and without $\mathcal{L}_M$, respectively. Comparing with Fig. 2 (c), the distributions of features in Fig. 2 (d) are more regularized with a closer center distance (as can be seen from the red center points). They reveal that $\mathcal{L}_M$ is beneficial to perform crossmodal refinement by reducing modality gap for better predictions.

The proposed UCRN emphasizes the importance of unimodal and crossmodal refinements. The contributions of different components are summarized in Table 6. We have the following observations: 1) URM greatly boosts the performance; 2) adding SQIL yields a better performance; 3) UCRN gains large improvement by adding the Multimodal JS divergence (MJS) and 4) CRM further adds values on top. The results substantiated all our assumptions that unimodal refinement has significant contributions and advantages to fusion and crossmodal refinement is efficacious in exploring modality-common information and reducing modality gap.

Model	avg F1	avg Acc	MAE	Corr
UCRN	**82.30**	**82.25**	**0.60**	**0.68**
UCRN w/o CRM	81.94	81.51	0.61	0.67
UCRN w/o SQIL	80.76	80.13	0.64	0.66
UCRN w/o MJS	80.30	80.49	0.63	0.65
UCRN w/o URM	65.83	64.79	0.78	0.25

Table 6: Ablation study in the 2-class sentiment task on CMU-MOSEI. UCRN includes URM, SQIL, CRM, and MJS components. Avg F1 and avg Acc are in percentages (%).

4.6 Size of Network Parameters

As listed in Table 7, UCRN is light-weight and can achieve competitive performance with much fewer parameters comparing with the several benchmark methods.

Method	Number of Parameters
TBJE	68,674,873
Mu-Net	21,499,212
MulT	1,071,211
UCRN(trimodal)	422,401
UCRN(bimodal)	211,681

Table 7: Network parameter comparison.

5 Conclusion

In this work, the Unimodal and Crossmodal Refinement Network (UCRN) is proposed for robust and efficient multimodal representation learning. We hypothesis that unimodal representation is better to be refined before crossmodal fusion, and it is beneficial to reduce modality gaps before crossmodal refinement. Following the line, the proposed network is designed with a unimodal refinement module, a multimodal JS divergence regularizer, self-quality improvement layers, and a crossmodal refinement module. The experimental results validated all our assumptions. In particular, the robustness experiments evinced high efficiency and validated that UCRN can handle the modality missing and noise issues. Experimental results showed UCRN achieves state-of-the-art results on multiple multimodal datasets.

Acknowledgements

This work is partially supported by the Ministry of Education, Singapore through Academic Research Fund Tier 1, RG21/19-(S). This work was partially accomplished during the author's internship at AARC. The author would like to thank the contributions from all the co-authors.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5651–5661.

Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 2018. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964. IEEE.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Paul Ekman, Wallace V Freisen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. In *Advances in Neural Information Processing Systems*, pages 473–479.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Runnan Li, Zhiyong Wu, Jia Jia, Sheng Zhao, and Helen Meng. 2019. Dilated residual network with multi-head self-attention for speech emotion recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6675–6679. IEEE.

Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 164–172.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.

Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*, pages 338–353. Springer.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *arXiv preprint arXiv:2002.08267*.

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 the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. 2019. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

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 the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

A Multiple Performance Evaluations

A.1 Accuracy Standard Deviation (acc-std)

In this section, the standard deviations of accuracies from Table 1 to Table 5 are summarized. In Table 8, 'acc2-std', 'acc7-std', and 'avg-emo-acc-std' represent the standards deviations of UCRN from 10 runs with different initializations for 2-class classification accuracy, 7-class classification accuracy, and average emotion classification accuracy, respectively. UCRN shows less in fluctuations on different tasks, which validates its performance stability.

Table	acc2-std	acc7-std	avg-emo-acc-std
Table 1	0.0049† 0.0036*	- 0.0042*	0.0096† 0.0063*
Table 2	0.0120	-	-
Table 3	0.0049	-	-
Table 4	0.0077	0.0135	0.0129
Table 5	0.0060	0.0045	0.0050

Table 8: Standard deviations of UCRN from Table 1 to Table 5. In Table 1, the results marked with † are from UCRN†, and * are from UCRN*

A.2 Relative Performance Improvement (RPI)

Table	Baselines	avg-acc2	avg-acc7	avg-emo-acc
Table 1	Mu-net TBJE	0.31† 7.94*	- 5.81*	0.50† 2.52*
Table 2	MulT Mu-Net	0.51 0.43	- -	- -
Table 3	MulT Mu-Net	1.33 0.31	- -	- -
Table 4	MulT Mu-net TBJE	2.72 1.63 0.12	- - 0.29	- 6.86 -
Table 5	MulT Mu-net TBJE	5.18 2.03 0.43	- - 10.96	- 15.14 -

Table 9: Relative Performance Improvement (RPI) of UCRN from Table 1 to Table 5. In Table 1, the results marked with † are from UCRN†, and * are from UCRN*

To show the significance of our results, the relative performance improvements are calculated to compare with the recent strong benchmark methods on different tasks. The metric is defined as $|our\ average\ result - baseline\ result|\ /our\ std.$ For example, to calculate the Acc7-RPI in Table 1 with TBJE, $X = |0.4684 - 0.444|\ /0.0042 = 5.81$.

As shown in Table 9, UCRN has higher performance improvement compared to all baseline methods and leads a large margin of RPIs on Acc-2 and Acc-7 with TBJE and MulT, and emo-acc with Mu-Net. On average, UCRN has 2.20, 3.74, and 4.01 relative performance improvement, comparing with MulT, Mu-Net, and TBJE, respectively.