BACN: Bi-direction Attention Capsule-based Network for Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Capsule-based network has currently identified its effectiveness in analyzing the heterogeneity issue of multimodal sentiment analysis. However, existing manners could only exploit the spatial relation between representation and output layer via down-top attention, which fails to effectively explore both inter-modality and intra-modality context. In this paper, during the preprocess period, we first present the multimodal dynamic enhanced module to facilitate the intra-modality context, which significantly boost the learning efficiency in dealing with multimodal heterogeneity issue. Furthermore, the bi-direction attention capsule-based network (BACN) is proposed to capture dynamic inter-modality context via the novel bi-direction dynamic routing mechanism. Specifically, BACN firstly highlights the static and low-level inter-modality context based on top-down attention. Then, the static multimodal context is transmitted to dynamic routing procedure, naturally allowing us to investigate dynamic and high-level inter-modality context. This indeed unleash the expressive power and provides the superior capability to bridge the modality gap among all the modalities. The experiments demonstrate that BACN can achieve state-of-the-art performance.

1 Introduction

Multimodal sentiment analysis has raised increasing interests in the artificial intelligence systems, where text, acoustic and visual modalities are popularly utilized to analyze the related research task(Ain et al., 2017; Rahman et al., 2020). The primary concern of multimodal analysis task is to learn a rich representation that better encapsulates two types of context: intra-modal and inter-modal context from multiple heterogeneous modalities. Indeed, the above context provide us the benefit to decrease the intra-modality and inter-modality redundancy simultaneously, allowing for effectively bridging the modality gaps of the heterogeneous modalities (Hazarika et al., 2020).

Recently, capsule-based networks have gained widespread attention for their significant performance in capturing the part-whole relationships among various modalities in computer vision and NLP(Lin et al., 2020), with the help of trainable viewpoint-invariant transformations. EF-Net (Wang et al., 2021) employed the standard capsule network to deal with the image presentation for exploring the spatial relation among distinct receptive areas of the image. In addition, McIntosh (McIntosh et al., 2020) proposed a capsule-based approach that introduced the novel visual-text routing mechanism for the integration of video and text modality. Nevertheless, the aforementioned techniques only attend to the spatial relation between representation layer and output layer via down-top attention. They indeed totally ignore the intra-modality context, and fail to effectively exploit the inter-modality context, leading to the great deterioration of task performance.

In this paper, during the preprocess period, the multimodal dynamic enhanced block is first proposed to explicitly facilitate the intra-modality context. This indeed effectively decrease the intra-modality redundancy of unimodality, and then significantly boost the learning efficiency in dealing with the multimodality heterogeneity issue. Furthermore, BACN is presented to exploit dynamic inter-modality context, using the novel bi-direction dynamic routing mechanism. Specifically, BACN firstly captures the static and low-level inter-modality context based on top-down attention. Then, the above static multimodal context is transmitted to the carefully designed multimodal dynamic routing process. This naturally gives learning model the strong ability to investigate dynamic and relatively high-level inter-modality context among multiple modalities. To the best of our knowledge, our model is the first dynamic multi-

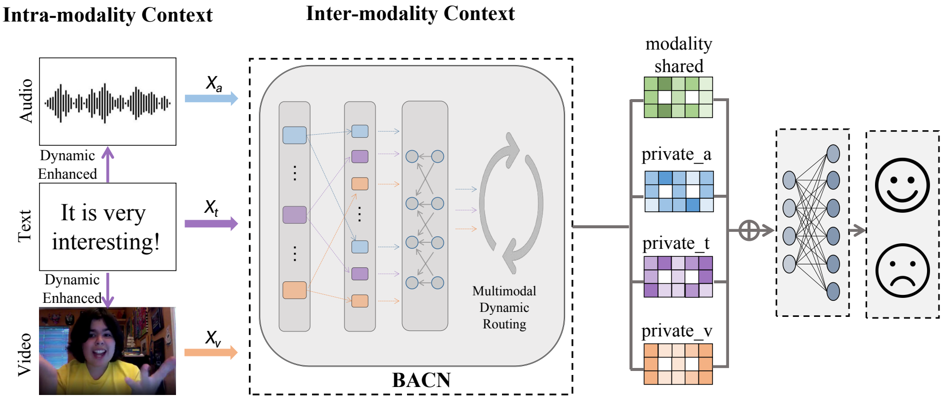


Figure 1: The overall architecture: Initially, during the preprocess period, the multimodal dynamic enhanced block is utilized to facilitate the intra-modality context of X_a (X_v), which significantly boost the learning efficiency in dealing with multimodality heterogeneity issue. Furthermore, BACN is proposed to exploit dynamic and high-level inter-modality context, allowing for effectively bridging the modality gaps of the heterogeneous modalities.

modal learning framework that supports the investigation of both the intra-modality and inter-modality task-related context. In addition, BACN has demonstrated the superiority on two multimodal learning benchmarks.

2 Related Work

The existing multimodal sentiment learning model consists of the following two leading lines:

Non Shared-Private Multimodal Learning Recently, LSTM and RNN based techniques have drawn a surge of interest in multimodal sentiment analysis for their excellence in exploiting the temporal correlation from the sequence data. For instance, *BC-LSTM* (Poria et al., 2017) proposed the bi-directional LSTM to highlight the contextual relationship among utterances. *RMFN* (Liang et al., 2018) utilized RNN to decompose the complex fusion process into several fusion sub-stages. Compared to the above models, attention-based frameworks have demonstrated the superiority in the long sequence presentation. *RAVEN* (Wang et al., 2019) applied the attention gating mechanism to compute the nonverbal shift vector. *MAG* (Rahman et al., 2020) introduced an attention gated memory to integrate the multimodal cues into the fusion context. Additionally, *MFN* (Zadeh et al., 2018a) leveraged the Delta-memory attention network to model the multimodal interactions. In addition, tensor-based models have raised increasing interests due to the high-dimension properties. *TFN* (Zadeh et al., 2017) employed the tensor manner to explicitly account for the unimodal, bimodal, and trimodal interactions. *LMF* (Liu et al., 2018) is the extension of *TFN*, which performs multimodal fusion process with designed modality-specific low-rank factors, significantly decreasing the computational complexity. However, the lack of minimizing the modality gap may limit their ability to effectively decrease the redundancy among modalities.

Shared-Private Multimodal Learning Broadly, the works of shared-private multimodal sentiment analysis could be categorized into the following three groups: 1) LSTM-based models: *MV-LSTM* (Rajagopalan et al., 2016) presented the multi-view LSTM block to explicitly model the view-private and view-shared interaction. Similarly, *MARN* (Zadeh et al., 2018b) applied the hybrid LSTM to store view-private and view-shared dynamics; 2) TopDown Attention based models: *MulT* (Tsai et al., 2019) proposed the cross-modal transformer to capture the static and low-level shared-representation. Similarly, *MCTN* (Pham et al., 2019) assigned the cyclic consistency loss to the standard Transformer, allowing for the joint representations. Different from *MulT* and *MCTN*, *MFM* (Tsai et al., 2018) factorized the joint distribution into shared information and modality-private message; 3) Correlation-based models focus on exploiting the modality-shared cues via Canonical Correlation Analysis (CCA) mechanism. *ICCN* (Sun et al., 2020) applied the deep CCA to retrieve the non-linear correlations among various modalities. Different from these models, *MISA* (Hazarika et al., 2020) employed the distribution similarity block to calculate similar portion across all modalities, and leveraged both shared and private information for sentiment prediction task. And, *Self-MM* (Yu et al., 2021) introduced unimodal subtasks to aid the modality-private representation learning. However, existing works have mainly focus on investigating the part-whole relation between low-level representation layer and high-level output layer. Indeed, they totally neglect the intra-modality context, and fail to effectively explore the inter-modality context, which raises a question on providing a deeper reasoning about multimodality heterogeneity issue.

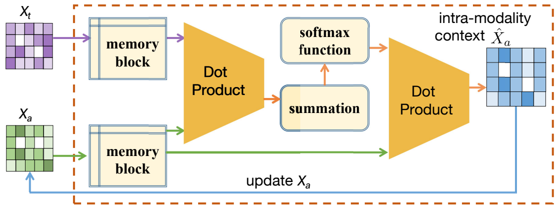


Figure 2: Multimodal dynamic enhanced block. Initially, X_a and X_t are leveraged to compute the bi-linear space via dot-product. Subsequently, the softmax function is utilized to exploit the context coefficients of X_a. Then, the coefficients are applied to measure the original X_a, leading to the more discriminative intra-modality context $\hat{X}_a$.

3 Methodology

As shown in Figure 1, the overall network consists of two essential components: 1) multimodal dynamic enhanced module that leveraged to facilitate the intra-modality context, and 2) BACN is further proposed to explore the inter-modality context.

3.1 Preliminaries

The two public sentiment benchmarks are composed of three modalities, audio, video and textual modality. The modality representation are represented as $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times d_a}$, $\boldsymbol{X_v} \in \mathbb{R}^{T_v \times d_v}$ and $\boldsymbol{X_t} \in \mathbb{R}^{T_t \times d_t}$, respectively. $T_i(i \in \{a, v, t\})$ refers to the number of utterances, and the feature dimension is denoted as $d_i(i \in \{a, v, t\})$. Note that, all modalities of original benchmarks have the same temporal dimension, i.e., $T_a = T_v = T_t$. Due to the properties of dot product, we adopt the linear function to analyze $\{X_a, X_v, X_t\}$ for retrieving the same feature dimension d_i, i.e., $d_a = d_v = d_t$.

3.2 Multimodal dynamic enhanced block

The multimodal dynamic enhanced block (Figure 2) is proposed to explicitly facilitate the intra-modality context of $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times d_a}$ ($\boldsymbol{X_v} \in \mathbb{R}^{T_v \times d_v}$), with the help of text modality($\boldsymbol{X_t} \in \mathbb{R}^{T_t \times d_t}$). Specifically, the presented block consists of M process heads, where each head comprises N adaptive iterations. Intuitively, the multi-head mechanism allows for extracting the intra-modality context with the multi-spect view, yields the comprehensive context. For the single-head case, the intra-modality context ${\boldsymbol{X_a}}_m^{[N_m]}$ of m-th head that associated with N_m iterations is formulated as follows:

$$
\begin{aligned}
&{\boldsymbol{X_a}}_m^{[N_m]} = f(\boldsymbol{X_a} \cdot \boldsymbol{X_t})\boldsymbol{X_a}, N_m = 1 \\
&{\boldsymbol{X_a}}_m^{[N_m]} = f(\sum_{i=1}^{N_m-1} {\boldsymbol{X_a}}_m^{[i]} \cdot \boldsymbol{X_t}) {\boldsymbol{X_a}}_m^{[N_m-1]}, N_m \geq 2, \qquad (1)
\end{aligned}
$$

where ‘f’ refers to the softmax function. During the first period of iteration, the dot-product operation is adopted to explicitly map the distinct modality into the bi-linear feature space $\boldsymbol{X_a} \cdot \boldsymbol{X_t}$. Subsequently, the softmax function is introduced to figure out how the utterances of the audio modality is influenced by the utterances in the text modality. Then, the context coefficients are applied to deal with the original audio modality, contributing to the more discriminative intra-modality context of audio. Due to the incorporation of the guidance from the more discriminative modality (text), the above process indeed provides us the strong ability to effectively investigate the intra-modality task-related context from auxiliary modality (audio and video).

On the basis of the first period of iteration, the next period of iteration attends to dynamically update the bi-linear space based on the output of the previous iteration. That is, the output data of previous iteration is leveraged to explore the new bi-linear space of next iteration, leading to the much more compact and robust bilinear space. Note that the process of $\boldsymbol{X_v}$ is similar to $\boldsymbol{X_a}$. Taking the single-head enhanced block as basis, the multi-head enhanced network is further established to collect the multiway intra-modality context message. Additionally, the convolution operation is introduced to analyze the multiway intra-modality context, which is able to further explore the latent interaction among distinct ${\boldsymbol{X_a}}_m^{[N_m]}$, leading to the much more compact task-related context $\hat{\boldsymbol{X}}_{\boldsymbol{a}}$.

$$
\hat{\boldsymbol{X}}_{\boldsymbol{a}} = Conv(concat({\boldsymbol{X_a}}_1^{[N_1]}, \cdots, {\boldsymbol{X_a}}_M^{[N_M]})) \qquad (2)
$$

Note that, during the preprocess period, we utilize the simple operation to analyze the more discriminative intra-modality context of auxiliary modality (audio and video). This indeed effectively decrease

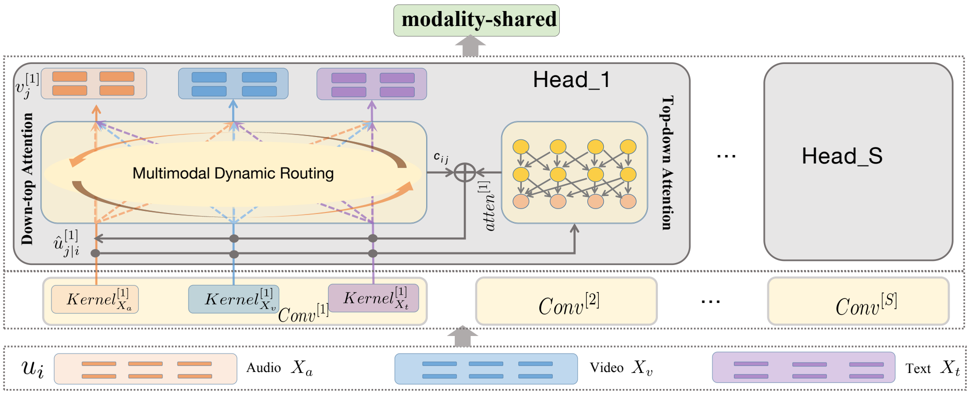


Figure 3: BACN: The u_i and v_j refers to the representation and modality-shared capsules, respectively. The static and low-level inter-modality context is firstly exploited based on top-down attention, and latently is transmitted to the carefully designed multimodal dynamic routing process. This naturally gives learning model the strong ability to investigate dynamic and relatively high-level inter-modality context among multiple heterogeneous modalities.

the intra-modality redundancy of unimodality, and then significantly boost the learning efficiency in dealing with the heterogeneity issue among multiple distinct modalities.

3.3 Bi-direction Attention Capsule-based Network

When the enhanced procedure is finished, BACN is further proposed to explore the inter-modality context. This indeed significantly boosts the learning efficiency and provides the superior capability to effectively investigate the inter-modality context among multiple more discriminative modalities.

As shown in Figure 3, BACN is mainly comprised of multimodal representation capsules $\{u_i\}_{i=1}^{N_u}$ and task-related capsules $\{v_j\}_{j=1}^{N_v}$, where N_u and N_v refer to the number of representation and task-related capsules respectively. Note that, $\{u_i\}$ are captured based on $\{X_a, X_v, X_t\}$ (Lin et al., 2020). In the conventional capsule network, each u_i is multiplied by a trainable transformation matrix W_{ij}, leading to the vote matrix $\hat{u}_{j|i}$ which stands for the projection of the representation u_i with respect to task-related capsule v_j, where $\hat{u}_{j|i} = u_i W_{ij}$.

Compared to the conventional capsule network, we replace the linear W_{ij} with the proposed convolution projection, resulting in new $\hat{u}_{j|i}$ consists of the desirable convolutional nonlinear properties. This allows for the more fine-grained projection procedure of representation capsule u_i with respect to task-related capsule v_j:

$$\begin{aligned}\hat{u}_{j|i} =& Conv(u_i, kernel_i)\\ =& sigmoid(\sum u_i * kernel_i + bias_i). \qquad (3)\end{aligned}$$

In addition, we extend the above single-head convolution projection design to the multi-head case associated with varying convolution kernels. Actually, the multi-head mechanism indeed allows for the multiway and comprehensive information flow between the representation capsule u_i and the task-related capsule v_j, where s refers to the specific convolution projection head:

$$\begin{aligned}\hat{u}_{j|i}^{[s]} &= Conv^{[s]}(u_i, kernel_i^{[s]})\\ &= sigmoid(\sum u_i * kernel_i^{[s]} + bias_i^{[s]}) \qquad (4)\end{aligned}$$

Note that, the down-top attention of capsule network could only analyze the part-whole (spatial) relation between representation capsules $\{u_i\}_{i=1}^{N_u}$ and task-related capsules $\{v_j\}_{j=1}^{N_v}$, with the help of dynamic routing coefficients c_{ij}. Actually, during the dynamic multimodal learning procedure, this fails to explicitly highlight the inter-modality context among distinct modality representations u_i, which shows its limitation in effectively reducing the inter-modality redundancy. Therefore, in this work, we first exploit the static and low-level inter-modality context among multiple modality representations u_i, based on top-down attention. Formally, the static inter-modality context $atten^{[s]}$ of the s-th head is defined as follows:

$$\begin{aligned}atten^{[s]} =& TopDownAttention([\hat{u}_{j|i_1}^{[s]}, ..., \hat{u}_{j|i_{N_u}}^{[s]}])\\ =& f(W_q[\{\hat{u}_{j|i}^{[s]}\}_{i=1}^{N_u}]W_k^T[\{\hat{u}_{j|i}^{[s]}\}_{i=1}^{N_u}]^T)W_v[\{\hat{u}_{j|i}^{[s]}\}_{i=1}^{N_u}], \qquad (5)\end{aligned}$$

where '[]' refers to the concatenation operation, 'f' indicates the softmax function, and $\{W_q, W_k, W_v\}$ are the transformation matrixes. Subsequently, the dynamic routing procedure with N_v iterations were conducted to explore the dynamic inter-modality context among multiple modalities. At each iteration, the dynamic coefficients $c_{ij}^{[s]}$ is leveraged to analyze the information flow between $\{u_i\}_{i=1}^{N_u}$ and $\{v_j\}_{j=1}^{N_v}$, which is calculated based on the temporary cumulant variable $b_{ij}^{[s]}$ that initialized as 0. That is to say, $c_{ij}^{[s]}$ could be utilized to measure how each task-related v_j is influenced by modality representations $\{u_i\}_{i=1}^{N_u}$. To reiterate, we attempt to leverage our BACN to exploit the modality-shared message among multiple modalities, thus the task-related v_j refers to the modality-shared message. The detailed procedure is formulated as follows:

$$\begin{aligned}\{c_{ij}^{[s]}\}_{j=1}^{N_v} &= Softmax(\{b_{ij}^{[s]}\}_{j=1}^{N_v}) \\ &= \frac{exp(b_{ij}^{[s]})}{\sum_{j=1}^{N_v} exp(b_{ij}^{[s]})}\end{aligned} \tag{6}$$

Then, task-related capsule $v_j^{[s]}$ is represented as the weighted sum of $\hat{u}_{j|i}^{[s]}$, with the help of corresponding $c_{ij}^{[s]}$ and the aforementioned static inter-modality context $atten^{[s]}$. It is important to note that, different from the conventional capsule-based network where $v_j^{[s]}$ only depends on $c_{ij}^{[s]}$ and $\hat{u}_{j|i}^{[s]}$, our model further transmit the static and low-level inter-modality context $atten^{[s]}$ to the carefully designed multimodal dynamic routing procedure. This indeed gives the learning model the strong ability to explore the dynamic and relatively high-level inter-modality context among multiple modalities. Essentially, the top-down attention mechanism simply investigate the static inter-modality context at once, leading to the relatively low-level context. On the contrast, we attempt to add the static inter-modality context $atten^{[s]}$ to the corresponding dynamic coefficients $c_{ij}^{[s]}$, allowing for the dynamic process of capturing the inter-modality context. Intuitively, the novel bi-direction dynamic coefficient $(c_{ij}^{[s]} + atten^{[s]})$ naturally allows us to dynamically modify inter-modality context during the novel bi-direction dynamic process that associated with multiple dynamic iterations, leading to the high-level inter-modality context.

$$v_j^{[s]} = \sum_i (c_{ij}^{[s]} + atten^{[s]}) \hat{u}_{j|i}^{[s]} \tag{7}$$

When the head is set to 2, each modality could compute two corresponding modality-shared messages $\{v_j^{[1]}, \ v_j^{[2]}\}$. Then, the above modality-shared messages could be further integrated into the unit modality-shared messages $\{shared_a, shared_v, shared_t\}$ via convolution operation. For instance, $shared_a = conv(concat(v_{j\_a}^{[1]}, v_{j\_a}^{[2]}), kernel_a)$. Then, all the modality-shared messages are further merged into the output $modality-shared$ via convolution operation : $modality-shared = conv(concat(shared_a, shared_v, shared_t), kernel)$.

As mentioned before, the convolution projection is leveraged to analyze the u_i, which allows for the convolutional nonlinear representation. Accordingly, we introduce the HingeLoss (Bailer et al., 2017) that attends to the analysis of nonlinear message for reducing the discrepancy among modality-shared messages:

$$\begin{aligned}SimilarityLoss &= \sum HingeLoss(shared_i, shared_j) \\ &= \sum max(0{,}1 - \|D(shared_i) - D(shared_j)\|_2),\end{aligned} \tag{8}$$

where $i, j \in \{a, v, t\}$, and $i \neq j$. Additionally, in our work, each modality-private message $private\_i$ is captured by the individual BACN, i.e., $private\_i = BACN(modality\_i)$. Then, following the constraint design of MISA, the difference loss is formulated as: $DifferenceLoss = \sum_{i \in \{a,v,t\}} \|shared_i^T private_i\|_F^2 + \sum_{i,j \in \{a,v,t\}} \|private_i^T private_j\|_F^2$.

4 Experiments Setups

4.1 Datasets

CMU-MOSI dataset (Zadeh et al., 2016) is comprised of 2199 utterance-video segments collected from 93 movie review videos of Youtube. Each utterance is manually annotated with the continuous sentimental label in the range of [-3, 3] from strong negative to strong positive. Additionally, the above dataset consists of 1284 training, 229 validation, and 686 testing samples. CMU-MOSEI dataset (Zadeh et al., 2018c) is the extension of CMU-MOSI associated with much more utterance segments. This version is composed of 22856 annotated utterances, and is split into the training, validation, and testing sets (16326, 1871, 4659).

4.2 Features and Evaluation Metrics

For CMU-MOSI and CMU-MOSEI, we adopt the same manner of MAG and MISA to extract the features of the specific modality. Specifically, the pre-trained BERT and XLNet are utilized to exploit the corresponding textual representations. Additionally, the following evaluation metrics are introduced to analyze the performance of the proposed model: mean absolute error (MAE), pearson correlation (Corr), binary accuracy (Acc-2), F-Score (F1). Essentially, two distinct manners are proposed to measure Acc-2 and F1. 1) In the work of (Zadeh et al., 2018b), the negative class is annotated with the label in the range of [-3, 0), while the range of non-negative class is [0, 3]. 2) On the contrast, in the work of (Tsai et al., 2019), the range of negative and positive class are [-3, 0) and (0, 3], respectively. The marker -/- is employed to distinguish the distinct strategies, where the left-side value refers to 1) and the right-side value stands for 2).

4.3 Comparisons

We introduced the non shared-private and shared-private multimodal learning models as the baselines. Non shared-private based: Bi-directional LSTM (BC-LSTM), RNN-based multistage fusion network (RMFN), Recurrent Attended Variation Embedding Network (RAVEN), Multimodal Adaptation Gate (MAG), Memory Fusion Network (MFN), Tensor Fusion Network (TFN), Low-rank Multimodal Fusion (LMF). Shared-private based: Multi-view LSTM (MV-LSTM), Multi-attention Recurrent Network (MARN), Multimodal Transformer (MulT), Multimodal Cyclic Translation Network (MCTN), Multimodal Factorization Model (MFM), Interaction Canonical Correlation Network (ICCN), Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis(MISA), Self-Supervised Multi-task Multimodal model (Self-MM).

4.4 Training Details

We perform the grid-search over the hyper-parameters to select the model with the best validation task loss. The range of essential hyper-parameters are summarized as follows: head [1, 6], iteration [1, 7], convolution kernel {3, 5, 7}.

5 Experiments results and analysis

5.1 Performance comparison with state-of-the-art models.

The performance of baselines, our proposed BACN and the ablation case BACN (Non-Enhanced) are illustrated in following tables. Note that, BACN (Non-Enhanced) refers to the case that BACN performs the multimodal learning task on the original modality data rather than the outputs of the enhanced block. The bottom rows in Table 1, Table 2 and Table 3 demonstrate the superiority and effectiveness of BACN. Particularly, on CMU-MOSEI benchmark, BACN exceeds the previous best Self-MM (bert) on the metric 'Corr' by a margin of 5.0%. Additionally, on CMU-MOSI dataset, BACN outperforms MISA (bert) on the metric 'Acc-7' with an improvement of 6.9%. The observations signify the necessity of exploiting the both the intra-modality and inter-modality task-related context. Essentially, we can observe that BACN obtains better results than the ablation case BACN (Non-Enhanced). This indicates that the enhanced block indeed effectively decrease the intra-modality redundancy of unimodality, which significantly boosts the learning efficiency in dealing with the multimodality heterogeneity issue.

Models	CMU-MOSI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	Acc-7(↑)
BC-LSTM	1.079	0.581	73.9/-	73.9/-	28.7
MV-LSTM	1.019	0.601	73.9/-	74.0/-	33.2
$RMFN^{\otimes}$	0.922	0.681	78.4/-	78.0/-	38.3
$RAVEN^{\otimes}$	0.915	0.691	78.0/-	76.6/-	33.2
MFN	0.965	0.632	77.4/-	77.3/-	34.1
MARN	0.968	0.625	77.1/-	77.0/-	34.7
TFN	0.970	0.633	73.9/-	73.4/-	32.1
LMF	0.912	0.668	76.4/-	75.7/-	32.8
MulT	0.871	0.698	-/83.0	-/82.8	40.0
$MCTN^{\otimes}$	0.909	0.676	79.3/-	79.1/-	35.6
$MFM^{\otimes}$	0.951	0.662	78.1/-	78.1/-	36.2
Capsule Network (Bert)	0.762	0.778	83/86	83.4/86.1	39.5
$TFN(Bert)^{\triangle}$	0.901	0.698	-/80.8	-/80.7	34.9
$LMF(Bert)^{\triangle}$	0.917	0.695	-/82.5	-/82.4	33.2
ICCN (Bert)	0.860	0.710	-/83.0	-/83.0	39.0
MISA (Bert)	0.783	0.761	81.8/83.4	81.7/83.6	42.3
MAG (Bert)	0.712	0.796	84.2/86.1	84.1/86.0	-
Self-MM (Bert)	0.713	0.798	84.0/85.98	84.42/85.95	-
ABCN (Non-Enhanced) (Bert)	0.684	0.824	86.0/88.4	85.9/88.4	47.8
ABCN (Bert)	**0.669**	**0.833**	**86.5/89.1**	**86.5/89.1**	**49.2**

Table 1: Performances of baselines and BACN based on BERT in CMU-MOSI benchmark. Note that (Bert) means the textual presentation is explored via BERT; ⊗ from (Tsai et al., 2019); △ from (Sun et al., 2020)

5.2 Effect of head and convolution kernel of BACN.

Note that, compared to the conventional capsule network, our proposed capsule-based framework (BACN) replace the linear transformation matrix with the presented multi-head convolution component. Therefore, we are interested to measure how varying heads and convolution kernel size affect the architecture performance. The head varies from 2 to 6, and each head is associated with a corresponding convolution kernel is of the same size (3×3, 5×5 or 7×7). In Figure 4, BACN is capable of receiving good results with respect to the head and kernel. Notably, kernel_3×3 based setting reaches

Models	CMU-MOSI			
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)
TFN	0.970	0.633	73.9/-	73.4/-
MARN	0.968	0.625	77.1/-	77.0/-
MFN	0.965	0.632	77.4/-	77.3/-
RMFN	0.922	0.681	78.4/-	78.0/-
MulT	0.871	0.698	-/83.0	-/82.8
Capsule Network (X)	0.75	0.799	83.7/85.9	83.8/85.9
$TFN(X)^{\diamondsuit}$	0.914	0.713	78.2/80.1	78.2/78.8
$MARN(X)^{\diamondsuit}$	0.921	0.707	78.3/79.5	78.8/79.6
$MFN(X)^{\diamondsuit}$	0.898	0.713	78.3/79.9	78.4/79.1
$RMFN(X)^{\diamondsuit}$	0.901	0.703	79.1/81.0	78.6/80.0
$MulT(X)^{\diamondsuit}$	0.849	0.738	87.9/84.4	80.4/83.1
MAG (X)	0.675	0.821	85.7/87.9	85.6/87.9
ABCN (Non-Enhanced) (X)	0.672	0.827	85.2/87.4	85.1/87.4
ABCN (X)	**0.661**	**0.836**	**86.6/88.8**	**86.5/88.8**

Table 2: Performances of baselines and BACN based on XLNet in CMU-MOSI benchmark. Note that (X) means the textual presentation is explored via XLNet; ♢ from (Rahman et al., 2020).

Models	CMU-MOSEI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1(↑)	Acc-7(↑)
$MFN^{\otimes}$	-	-	76.0/-	76.0/-	-
$MV-LSTM^{\otimes}$	-	-	76.4/-	76.4/-	-
RAVEN	0.614	0.662	79.1/-	79.5/-	50.0
MCTN	0.609	0.670	79.8/-	80.6/-	49.6
MulT	0.580	0.703	-/82.5	-/82.3	51.8
Capsule Network (Bert)	0.581	0.80	83.8/86.4	84/86.3	48.6
$TFN(Bert)^{\triangle}$	0.593	0.700	-/82.5	-/82.1	50.2
$LMF(Bert)^{\triangle}$	0.623	0.677	-/82.0	-/82.1	48.0
$MFM(Bert)^{\triangle}$	0.568	0.717	-/84.4	-/84.3	51.3
ICCN (Bert)	0.565	0.713	-/84.2	-/84.2	51.6
MISA (Bert)	0.555	0.756	83.6/85.5	83.8/85.3	52.2
Self-MM (Bert)	0.530	0.765	83.79/85.23	83.74/85.3	-
ABCN (Non-Enhanced) (Bert)	0.563	0.806	85.3/86.9	85.2/86.8	49.9
ABCN (Bert)	**0.551**	**0.815**	**86.3/87.1**	**86.1/87.1**	51.3

Table 3: Performances of baselines and BACN based on BERT in CMU-MOSEI benchmark. Note that (Bert) means the textual presentation is explored via BERT; ⊗ from (Zadeh et al., 2018c); △ from (Sun et al., 2020).

the peak value at head 4, and kernel_5×5 based setting maximizes prediction performance at head 3. This implies that multi-head strategy is able to give each head the strong ability to exploit the essential and comprehensive sentimental polarity, allowing for the multi-level multimodal message. Moreover, the setting which consists of too many heads may contribute to similar multimodal presentation pattern within the same feature map, leading to large information redundancy. On the contrast, the setting which is comprised of too few heads may fail to effectively explore the sufficient multimodal interactions. It is interesting to find that, compared to the kernel_3×3 and kernel_5×5 based setting, kernel_7×7 based setting receives the best performance at head 5. Actually, kernel_7×7 attempts to process the multimodal fusion procedure within the large receptive field, which may lead to the lack of fine-grained and local intercorrelations among multiple modalities to some extend.

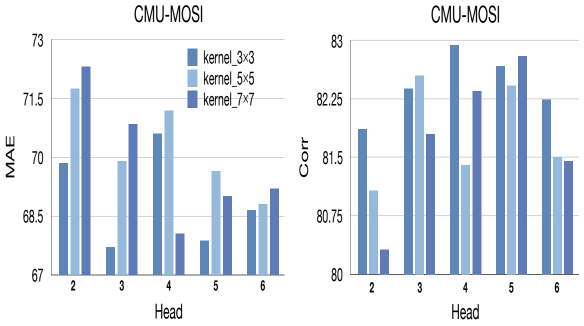


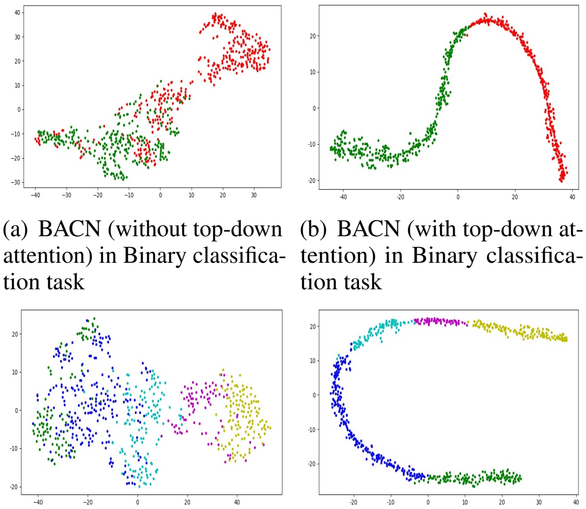
Figure 4: Results of effect of head and convolution kernel on CMU-MOSI.

5.3 Effect of top-down attention of BACN.

In this work, compared to the conventional capsule network, BACN first exploit the static and low-level inter-modality context via the top-down attention. Therefore, we attempt to investigate how top-down attention affects the classification task. Specifically, t-SNE method is utilized to provide the corresponding visualization of the multimodal fusion representations learned by BACN. For the binary classification task, the red points refer to the positive sentiment, and the green points indicate the negative sentiment. For the multi-classification task, the color of the points depends on the corresponding annotated sentimental labels. In Figure 5, we can observe that the multimodal fusion message becomes increasingly separable when BACN is associated with the top-down attention mechanism. Actually, the top-down attention mechanism is able to naturally benefit the down-top attention based network to explicitly explore the dynamic and relatively high-level inter-modality context message, leading to the significant improvement of discriminative efficiency and expressive capability.

5.4 Effect of the head of multimodal dynamic enhanced block.

In this work, the multimodal dynamic enhanced block is proposed to explicitly facilitate the intra-modality context. Specifically, the proposed enhanced block is comprised of M process heads. Therefore, we are interested to investigate how distinct heads affect the task performance. The head varies from 1 to 6. As shown in Figure 6, our proposed model is capable of obtaining fairly good performance with respect to the enhanced heads. It is important to observe that, our model reaches the peak value at the head 2 for the case of CMU-MOSI (Bert). As to the CMU-MOSEI (Bert), we



(a) BACN (without top-down attention) in Binary classification task

(b) BACN (with top-down attention) in Binary classification task

(c) BACN (without top-down attention) in Multi-classification task

(d) BACN (with top-down attention) in Multi-classification task

Figure 5: t-SNE visualization of the multimodal fusion presentation learned by BACN on CMU-MOSI.

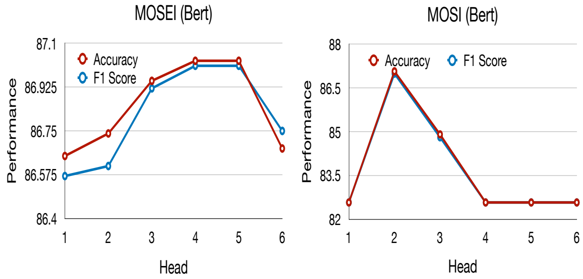


Figure 6: Effect of the head of multimodal dynamic enhanced block on CMU-MOSI and MOSEI.

can observe that the relatively higher performance is received at head 4. Indeed, the multi-head mechanism allows for exploiting the intra-modality context with the multi-spect view, yields comprehensive context. Accordingly, the proposed multi-head enhanced strategy provides us the benefit of further boosting the expressive efficiency and capability. Additionally, the too-simple enhanced bock which is comprised of too few heads (e.g., 1 head) may fail to effectively discover the comprehensive intra-modality context. And, the too complex enhanced block that consists of too many heads may provide large similar intra-modality context, leading to the information redundancy and the greater performance drop.

5.5 Effect of the dynamic iteration of multimodal dynamic enhanced block.

As mentioned before, the proposed multimodal dynamic enhanced block is comprised of M process heads, and each head consists of N adaptive iterations. In this part, we attempt to analyze how various adaptive iterations affect the model performance. The number of adaptive iterations ranges from 1 to 7. For simplicity, we only perform the relative ablation study on the 1-head setting. As shown in Figure 7, our proposed model can obtain fairly good performance with respect to the adaptive iterations. It is interesting to find that, our model maximizes the task performance at the adaptive iteration 4 for the case of CMU-MOSI (bert). For CMU-MOSI (XLNet), we can observe that the relatively better performance is received at the adaptive iteration 3. Intuitively, each adaptive iteration attends to exploit the intra-modality context based on the more discriminative modality (text). On the basis of single adaptive iteration, the stacked iterations focus on dynamically update or modify the intra-modality context. This indeed effectively reduces the intra-modality redundancy of unimodality, and then significantly boost the learning efficiency in dealing with the heterogeneity issue among multiple distinct modalities.

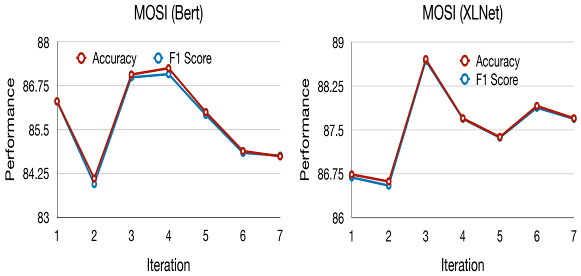


Figure 7: Effect of the dynamic iteration of multimodal dynamic enhanced block on CMU-MOSI.

6 Conclusion

In this paper, we first propose a simple multimodal enhanced module to facilitate the intra-modality context, which indeed effectively decrease the intra-modality redundancy of unimodality. Then, a novel bi-direction multimodal dynamic routing mechanism is presented to explicitly exploit dynamic and high-level inter-modality context. This indeed provides us the benefit to significantly boost the learning efficiency in dealing with the heterogeneity issue among multiple distinct modalities. To the best of our knowledge, our model is the first dynamic multimodal learning network that supports the investigation of both the intra-modality and inter-modality task-related context.

References

Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A Rehman. 2017. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, 8(6):424.

Christian Bailer, Kiran Varanasi, and Didier Stricker. 2017. Cnn-based patch matching for optical flow with thresholded hinge embedding loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3250–3259.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2020. Visual-textual capsule routing for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9942–9951.

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, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

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.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

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.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Jiaqian Wang, Donghong Gu, Chi Yang, Yun Xue, Zhengxin Song, Haoliang Zhao, and Luwei Xiao. 2021. Targeted aspect based multimodal sentiment analysis: an attention capsule extraction and multi-head fusion network. *arXiv preprint arXiv:2103.07659*.

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 the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *CoRR*, abs/2102.04830.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

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, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

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.