# MMT: Multi-Way Multi-Modal Transformer for Multimodal Learning

**Jiajia Tang**[1*], **Kang Li**[1*], **Ming Hou**[2], **Xuanyu Jin**[1], **Wanzeng Kong**[1†], **Yu Ding**[3] and **Qibin Zhao**[2]

[1]Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, School of Computer Science and Technology, Hangzhou Dianzi University, China

[2]RIKEN Center for Advanced Intelligence Project (AIP), Japan

[3]Virtual Human Group, Netease Fuxi AI Lab

{hdutangjiajia, jxuanyu599}@163.com, {likang_bro, kongwanzeng}@hdu.edu.cn, {ming.hou, qibin.zhao}@riken.jp, dingyu01@corp.netease.com

## Abstract

The heart of multimodal learning research lies the challenge of effectively exploiting fusion representations among multiple modalities. However, existing two-way cross-modality unidirectional attention could only exploit the intermodal interactions from one source to one target modality. This indeed fails to unleash the complete expressive power of multimodal fusion with restricted number of modalities and fixed interactive direction. In this work, the multiway multimodal transformer (MMT) is proposed to simultaneously explore multiway multimodal intercorrelations for each modality via single block rather than multiple stacked cross-modality blocks. The core idea of MMT is the multiway multimodal attention, where the multiple modalities are leveraged to compute the multiway attention tensor. This naturally benefits us to exploit comprehensive many-to-many multimodal interactive paths. Specifically, the multiway tensor is comprised of multiple interconnected modality-aware core tensors that consist of the intramodal interactions. Additionally, the tensor contraction operation is utilized to investigate intermodal dependencies between distinct core tensors. Essentially, our tensor-based multiway structure allows for easily extending MMT to the case associated with an arbitrary number of modalities. Taking MMT as the basis, the hierarchical network is further established to recursively transmit the low-level multiway multimodal interactions to high-level ones. The experiments demonstrate that MMT can achieve state-of-the-art or comparable performance.

## 1 Introduction

Multimodal learning is a very actively growing field of research, which has witnessed many significant advances in sentiment analysis and speaker personality recognition [Ain *et al.*, 2017]. Indeed, the multimodal signals such as text modality, visual modality, and acoustic modality consist of complementarity and consistency. Consequently, the heart of multimodal learning lies the challenge of effectively accounting for the both intra-modality and inter-modality intercorrelations among multiple modalities.

More recently, with the advent of self-attention mechanism, transformer-based fusion frameworks [Soleymani *et al.*, 2017] have been extensively employed for exploiting the long-range dependencies of modalities. For instance, MISA [Hazarika *et al.*, 2020] and MAG [Rahman *et al.*, 2020] leverage self-attention to retrieve the contextual embeddings of textual modality. Note that, the self-attention focuses on highlighting the intra-modality dependencies within the individual modality, leading to a lack of explicit interaction among different modalities. On the basis of conventional self-attention, MulT [Tsai *et al.*, 2019] proposes a novel cross-modality two-way attention component, which benefits the model to directly explore the explicit cross-modality intercorrelations within two-way space. However, the unidirectional cross-modality attention could only exploit the intermodality interactions from one source to one target modality. Therefore, the above architectures have shown the limitation in offering the much more representative capability with the fixed interactive direction and restricted number of involved modalities, e.g., only up to two modalities. This indeed overlooks the complex and essential global interactions among more modalities, which results in the information loss and the deterioration of prediction. Essentially, their model attempt to sequentially stack distinct two-way cross-modality attention blocks into the hierarchical one for the multimodal learning task, ie., $(text \rightarrow visual) \rightarrow acoustic$. Intuitevely, this sequential one-to-one procedure $\{Modality_i\} \rightarrow \{Modality_j\}$ would lead to the significant increase of cost in computation and memory.

To overcome the above underlying research issues, we propose the multi-way multi-modal Transformer (MMT), a method that extends the standard Transformer framework to analyze multiple modalities simultaneously. The core idea of MMT is the tensor-based multiway multimodal attention, which provides an M-dimension

*Equal contribution

†Corresponding author: Wanzeng Kong

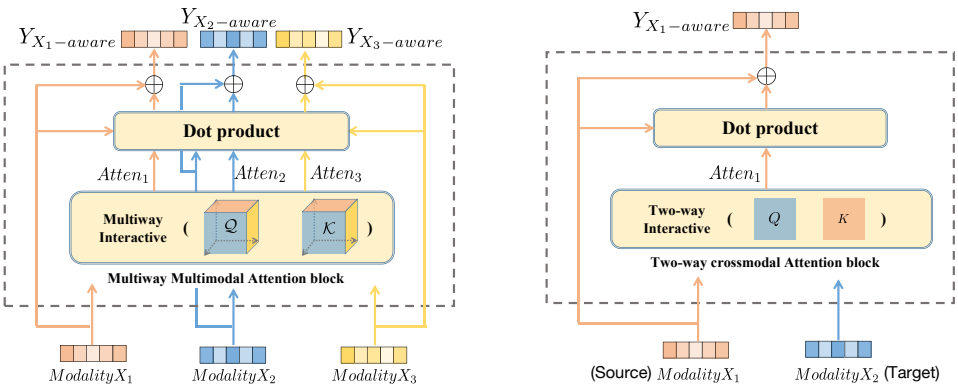


Figure 1: Comparison of MMT with existing crossmodal attention model. In our model, multiway multimodal attention is proposed to exploit multiple modality-aware multimodal intercorrelations simultaneously via the single block. Note that, compared to the two-way attention model, MMT is able to provide a multiway attention space that consists of many-to-many interactive paths, which significantly boosting the expressive capability and efficiency of the learning model.

attention space that consists of many-to-many multimodal interactive paths $\{Modality_1, \cdots, Modality_M\} \to \{Modality_1, \cdots, Modality_M\}$, where M indicates the modality number. This naturally contributes to both intra-modality and complex inter-modality interactions. Thanks to the low-rank tensor fashion, the proposed module has the superior capability to simultaneously explore the multiway multimodal attention space for each modality within the single attention block rather than multiple stacked cross-modality blocks. Indeed, the presented fashion allows for all the potential and comprehensive multiway interactions, as well as a much lighter model. It is important to note that, compared to the existing two-way attention module, the tensor-based multiway structure naturally provides us the great flexibility to extend MMT to the case associated with an arbitrary number of modalities. Intuitively, MMT is able to scale linearly in the number of modalities. Subsequently, taking MMT as the basis, the hierarchical architecture is established to transmit the low-level multiway multimodal intercorrelations to the high-level sophisticated presentation with the recursive fashion.

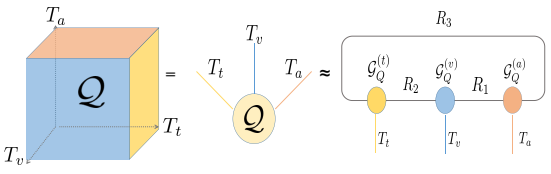


Figure 2: For the tensor-ring based $\mathcal{Q}$, the corresponding entries are expressed as $q_{t_t,t_v,t_a} = tr(G_Q^{(t)}(:,t_t,:)G_Q^{(v)}(:,t_v,:)G_Q^{(a)}(:,t_a,:))$.

## 2 Related Work

### 2.1 Non-attention based Model

RMFN [Liang *et al.*, 2018] introduced the RNN to process the data with the multistage fusion method. BC-LSTM [Poria *et al.*, 2017] presented a bi-directional LSTM to exploit the temporal dependencies. MV-LSTM [Rajagopalan *et al.*, 2016] and Self-MM [Yu *et al.*, 2021] utilized the multi-view (multi-task) strategy to compute sentiment information. Analogously, ICCN [Sun *et al.*, 2020], DF [Nojavanasghari *et al.*, 2016] and MISA attempt to capture the correlations across multiple modalities via similarity bock. Moreover, tensor-based works such as TFN [Zadeh *et al.*, 2017] and LMF [Liu *et al.*, 2018] proposed a multi-dimension fusion fashion to explicitly highlight multimodal intercorrelations among multiple messages.Nevertheless, the above frameworks fail to effectively draw the long-range temporal dependencies along the sequence.

### 2.2 Attention-based Model

Compared to the above models, the attention-based network such as MFM [Tsai *et al.*, 2018] is able to model long-range dependencies of the modality. RAVEN [Wang *et al.*, 2019] proposed a multimodal attention gating block. Similarly, MAG proposed the multimodal attention gating component. MCTN [Pham *et al.*, 2019] utilized the cyclic translation network to investigate multimodal representations. MFN [Zadeh *et al.*, 2018a] leveraged the Delta-memory attention to analyze the private modality portion. On the contrast, MARN [Zadeh *et al.*, 2018b] employed multimodal attention to highlight cross-modality dynamics. MEMI [Wu *et al.*, 2020] introduced the cross-modality attention to measure the two-way interactions. Additionally, MulT presented the two-way cross-modality attention to exploit the inter-modality unidirectional interactions. However, the above manners fail to exploit all the potential and comprehensive multimodal interactions, due to the fixed interactive direction and restricted number of involved modalities (only up to two modalities). Essentially, the existing models attempt to stack multiple cross-modality blocks into the hierarchical one for realizing the multimodal learning task, thus suffering from the significant increase of computational complexity and memory.

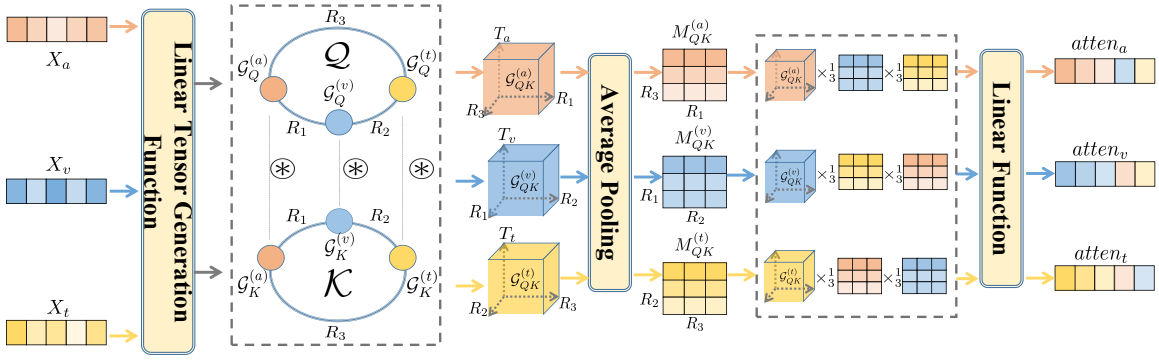


Figure 3: Multiway Multimodal Attention Block: Given the modality presentations $X_i$, the proposed attention block attempts to exploit the multiway attention space $\mathcal{QK}$ that consists of many-to-many interactive paths $\{X_1, \cdots, X_I\} \rightarrow \{X_1, \cdots, X_I\}$. This indeed brings forth the much more comprehensive and sophisticated multimodal intercorrelations, as well as the boost of expressive capability.

# 3 Methodology

## 3.1 Preliminaries

In this part, we attempt to introduce the notation of the tensor network utilized in our model, and present in detail the conventional self-attention and cross-modality attention mechanism.

### Tensor Network

Multi-dimensional array can be represented by *tensor*. Specifically, an $N$th-order tensor $\mathcal{S} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ has N dimensions, and the specific entry of $\mathcal{S}$ is denoted by $s_{i_1,i_2,\cdots,i_N} = \mathcal{S}(i_1, i_2, \cdots, i_N)$, where $i_N \in \{1, 2, \cdots, I_N\}$. Additionally, the tensor network - Tensor Ring can be further employed to decompose the higher-order tensor $\mathcal{S}$ into the low-rank case. That is to say, a set of interconnected lower-rank tensors (core tensors) $[\![\mathcal{G}^{(1)}, \cdots, \mathcal{G}^{(N)}]\!]$ could be utilized to represent $\mathcal{S}$. Notation $[\![]\!]$ is a simplified version of tensor ring decomposition. Note that, the hadamard product denoted as $\circledast$ and the mode-1 Khatri-Rao product denoted as $\odot_1$ are the essential operators in tensor analysis. Given two tensors $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ and $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, the hadamard product yieds a 3-order tensor $\mathcal{C} = \mathcal{A} \circledast \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ with entries $\mathcal{C}(i_1, i_2, i_3) = \mathcal{A}(i_1, i_2, i_3)\mathcal{B}(i_1, i_2, i_3)$. Given two matrixes $A \in \mathbb{R}^{N \times P_1}$ and $B \in \mathbb{R}^{N \times P_2}$, the mode-1 Khatri-Rao product yieds a matrix $C = A \odot_1 B = \in \mathbb{R}^{N \times P_1 P_2}$.

### Self-attention and Cross-modality Attention

At the heart of Transformer is the self-attention mechanism, which attends to the long-range dependencies within the uni-modality sequence. More specifically, Transformer is comprised of query, key and value matrix denoted as Q, K, V respectively. Considering the modalities set {Audio, Video, Text}, the utterance-level presentation are represented as $X_a \in \mathbb{R}^{T_a \times d_a}$, $X_v \in \mathbb{R}^{T_v \times d_v}$, and $X_t \in \mathbb{R}^{T_t \times d_t}$, respectively. In practice, $T_i$ is utilized to denote sequence length, and $d_i$ refers to the feature dimension. The corresponding audio-based self-attention procedure can be written as:$Attention(Q_a, K_a, V_a) = softmax(\frac{Q_a K_a^T}{\sqrt{d_k}})V_a$, which only focus on the intra-modality correlations from the uni-modality feature space.

On the basis of self-attention mechanism, MulT proposed the cross-modality attention mechanism, which performs the unidirectional cross-modal adaption within the two-way fusion space that consists of two corresponding modalities. For instance, the adaption $text(t) \rightarrow audio(a)$ can be written as $Attention(Q_a, K_t, V_t) = softmax(\frac{Q_a K_t^T}{\sqrt{d_k}})V_t$, which attends to the inter-modality correlations. Note that, the two-way cross-modality attention requires 4 transformer blocks $\{MulT_{t\rightarrow a},\ MulT_{a\rightarrow t},\ MulT_{a\rightarrow a},\ MulT_{t\rightarrow t}\}$ to obtain the cross-modality dependencies that consist of both intra-modality and inter-modality interactions. More importantly, the extension of MulT to M modalities would require $M^2$ transformer blocks, which results in the significant increase of computations and storage.

## 3.2 Multi-way Multi-modal Transformer

In this paper, tensor-based multi-way multi-modal attention is proposed to produce the rich representations for each modality by simultaneously paying attention to all the potential multiway interactions. Overall, our proposed MMT could accepts multiple inputs $\{X_a, X_v, X_t\}$, and produces the modality-aware multiway multimodal fusion outputs $\{Y_{a-aware}, Y_{v-aware}, Y_{t-aware}\}$.

Given modality presentation $\{X_t, X_v, X_a\}$, we first adopt a tensor-ring based generation function for retrieving the multiway multimodal query tensor $\mathcal{Q} = [\![\mathcal{G}_Q^{(t)}, \mathcal{G}_Q^{(v)}, \mathcal{G}_Q^{(a)}]\!] \in \mathbb{R}^{T_a \times T_v \times T_t}$ and key tensor $\mathcal{K} = [\![\mathcal{G}_K^{(t)}, \mathcal{G}_K^{(v)}, \mathcal{G}_K^{(a)}]\!] \in \mathbb{R}^{T_a \times T_v \times T_t}$. Note that, the tensor ring format naturally provides us the benefit of efficiently measuring the attention score on the low-rank core tensors ($\mathcal{G}_Q^{(i)} \in \mathbb{R}^{T_i \times R_w \times R_s}$, $\mathcal{G}_K^{(i)} \in \mathbb{R}^{T_i \times R_w \times R_s}$) rather than original large tensors ($\mathcal{Q}$, $\mathcal{K}$), contributing to the great decrease of model storage and complexity, where the index i $\in \{a, v, t\}$. Note that, $R_w$ and $R_s$ refer to the tensor-ring rank, the index w and s $\in \{1, 2, 3\}$, and w $\neq$ s. The above function can be formulated as:

$$\mathcal{G}_Q^{(i)} = reshape((X_i W_{Q_i}^{(1)}) \odot_1 (X_i W_{Q_i}^{(2)})) \tag{1}$$

where the linear transformation matrixes $\{W_{Q_i}^{(1)} \in \mathbb{R}^{d_i \times R_w}$,

$W_{Q_i}^{(2)} \in \mathbb{R}^{d_i \times R_s}\}$ are utilized to account for the low-rank projection of input. Additionally, the mode-1 Khatri-Rao product $\odot_1$ is introduced to incorporate the low-rank projections into the modality-aware core tensor $\mathcal{G}_Q^{(i)}$. Similarly, the key tensor $\mathcal{K}$ is received based on the same generation procedure.

Based on the above query tensor $\mathcal{Q}$ and key tensor $\mathcal{K}$ with the tensor ring format, the multiway multimodal attention procedure can be presented as follows:

$$MMT(\mathcal{Q},\mathcal{K}) = MMT([\![\mathcal{G}_Q^{(t)},\mathcal{G}_Q^{(v)},\mathcal{G}_Q^{(a)}]\!],[\![\mathcal{G}_K^{(t)},\mathcal{G}_K^{(v)},\mathcal{G}_K^{(a)}]\!]). \tag{2}$$

Then, the Hadamard product $\circledast$ is introduced to analyze the similarity between $\mathcal{G}_Q^{(i)}$ and $\mathcal{G}_K^{(i)}$, which allows for exploiting modality-based attention spaces $\mathcal{G}_{QK}^{(i)} \in \mathbb{R}^{T_i \times R_w \times R_s}$ in parallel. Indeed, above process is analogous to self-attention process ($QK^T$) in the standard Transformer framework. This naturally results in the multiway multimodal attention tensor $\mathcal{QK} = [\![\mathcal{G}_{QK}^{(t)},\mathcal{G}_{QK}^{(v)},\mathcal{G}_{QK}^{(a)}]\!] \in \mathbb{R}^{T_a \times T_v \times T_t}$. The attention core tensor $\mathcal{G}_{QK}^{(i)}$ are presented as follows:

$$\mathcal{G}_{QK}^{(i)} = \mathcal{G}_Q^{(i)} \circledast \mathcal{G}_K^{(i)} \tag{3}$$

Subsequently, the corresponding attention pooling matrixes $M_{QK}^{(i)} \in \mathbb{R}^{R_w \times R_s}$ are captured by averaging the $\mathcal{G}_{QK}^{(i)}$ along the temporal dimension ($T_i$), where each element of $M_{QK}^{(i)}$ is comprised of the intra-modality temporal dependencies. Notably, the $M_{QK}^{(i)}$ are then transmitted to the next process for integrating the distinct intra-modality attribution into the multi-way multi-modality attention message.

$$M_{QK}^{(i)} = average(\mathcal{G}_{QK}^{(i)}(:,j,p)), 1 \le j \le R_w, 1 \le p \le R_s \tag{4}$$

Given the modality-based attention pooling matrixes $M_{QK}^{(i)}$, we now attempt to explore the modality-aware multiway multimodal attention $atten_t$, $atten_v$ and $atten_a$. The detailed process is formulated as:

$$atten_i = linear(\mathcal{G}_{QK}^{(i)} \times_3^1 M_{QK}^{(j_1)} \times_3^1 \cdots \times_3^1 M_{QK}^{(j_{M-1})}) \in \mathbb{R}^{T_i \times d_i} \tag{5}$$

where $atten_i$ refer to the modality-aware multiway multimodal attention matrix, 'M' is the modality number, $i \in \{a,v,t\}$, $j_m \in \{a,v,t\}$, and $j_m \neq i$. Moreover, operation $\times_n^m$ refers to the mode-($_n^m$) product (tensor contraction operation). More specifically, the mode -($_3^1$) product of tensor $\mathcal{G}_{QK}^{(t)}$ and matrix $M_{QK}^{(a)}$, with the common mode $R_3$, yields a 3-order cross-modality attention tensor is of $T_t \times R_2 \times R_1$. Actually, the above process allows us to efficiently compute the cross-modality attention space, e.g., $(\mathcal{G}_{QK}^{(t)} \times_3^1 M_{QK}^{(a)})$, which effectively highlights the intercorrelations between text and audio modality. Compared to the conventional cross-modality two-way attention that only focuses on the inter-modality interactive path, our method simultaneously attends to the both intra-modality and complex inter-modality interactive paths. Then, the mode -($_3^1$) product of the above 3-order tensor $(\mathcal{G}_{QK}^{(t)} \times_3^1 M_{QK}^{(a)})$ and matrix $M_{QK}^{(v)}$, with the common mode $R_1$, yields a 3-order tensor is of $T_t \times R_2 \times R_2$. Intuitively, above procedure allows the cross-modality tensor further absorb the contribution from video modality, contributing to the comprehensive many-to-many multimodal interactive paths $\{t,a,v\} \rightarrow \{t,a,v\}$. Note that, $\{t,a,v\} \rightarrow \{t,a,v\}$ is comprised of the comprehensive many-to-many multimodal interactive paths, which could provide intra-modality and complex inter-modality interactions simultaneously. More specifically, $X_i \rightarrow X_i$ refer to the intra-modality interactions, and $X_i \rightarrow X_j$ refer to the inter-modality interactions ($i \neq j$), where $i,j \in \{t,a,v\}$. Then, the linear projection function is further applied to retrieve the text-aware multiway multimodality attention matrix $atten_t$.

In an addition, the modality-aware multiway multimodality presentation is achieved with the help of corresponding multiway attention matrix and a parameter $a$. Note that, $a$ is introduced to blend the contribution of modality-aware multiway multimodality message and the original modality information, leading to the much more comprehensive multimodal representation.

$$Y_{i-aware} = atten_i X_i + aX_i \in \mathbb{R}^{T_i \times d_i} \tag{6}$$

Note that, the tensor-ring function allows us to exploit inter-modality attention from low-rank small core tensors rather than original Mth-order large tensor, where M is the number of modality. This indeed significantly reduce the complexity and storage. If tensor-ring function is done after getting attention, the number of parameters of Mth-order tensor grows exponentially with 'M', i.e., parameters = $2 \times T_1 \times T_2 \times, \cdots, \times T_M$.

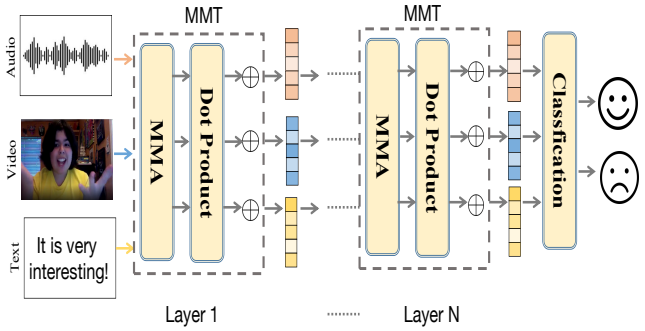


Figure 4: The hierarchical framework associated with N stacked MMTs. The proposed network allows us to transmit the low-level multiway multimodal intercorrelations to the high-level compact presentation with the recursive manner.

### 3.3 Hierarchical Architecture

On the basis of MMT, a hierarchical architecture was proposed for transmitting the low-level multiway multimodal intercorrelations to the much more comprehensive and expressive high-level intercorrelations with the recursive fashion. As illustrated in Figure 4, the proposed framework is comprised of N stacked MMT blocks. Specifically, the first MMT accepts the inputs $\{X_a, X_v, X_t\}$ and computes the modality-aware multiway multimodal presentation outputs set $\{Y_{a-aware}, Y_{v-aware}, Y_{t-aware}\}$. Subsequently, the next MMT takes the outputs of the previous MMT as the inputs, and latently performs the similar multiway multimodal attention procedure. As a result, the much more comprehensive and expressive multimodal intercorrelations can be identified and transmitted to the high-level layer via hierarchical structure. The overall recursive procedure is formulated as follows:

$$
\begin{aligned}
&[Y^{(1)}_{a-aware}, Y^{(1)}_{v-aware}, Y^{(1)}_{t-aware}] = MMT^{(1)}(X_a, X_v, X_t) \\
&[Y^{(i+1)}_{a-aware}, Y^{(i+1)}_{v-aware}, Y^{(i+1)}_{t-aware}] \\
&= MMT^{(i+1)}(Y^{(i)}_{a-aware}, Y^{(i)}_{v-aware}, Y^{(i)}_{t-aware}), i \in [1, N] \quad (7)
\end{aligned}
$$

Note that, due to the tensor-based multiway attention strategy, the $(i+1)-th$ layer of the proposed hierarchical architecture is able to accept the inputs $\{Y^{(i)}_{a-aware}, Y^{(i)}_{v-aware}, Y^{(i)}_{t-aware}\}$ from the previous $i-th$ layer recursively. That is to say, the presented multiway scheme indeed provides great expressive capability and flexibility of the learning model, giving each layer the strong ability to efficiently perform similar multiway multimodal attention based on the previous multiple outputs.

## 4 Experiments Setups

### 4.1 Datasets

The public sentiment benchmark CMU-MOSI [Zadeh *et al.*, 2016] is comprised of the aligned and preprocessed audio, video and text modality. CMU-MOSI consists of 2199 opinion video clips. Each clip is annotated with the corresponding sentiment intensity in the range of [-3, +3], spanning from the strongly negative to the strongly positive sentiments. The 2199 clips are spilt into 1284 train samples, 229 validation samples, and 686 test samples. The POM dataset [Park *et al.*, 2014] contains 903 movie opinion videos, which attends to the speaker trait recognition. Each video is annotated for following speaker traits with the strength score [1, 5] or [1, 7] : confident (con), passionate (pas), dominant (dom), vivid (viv), expertise (exp), entertaining (ent), and etc. The division of the train, validation and test sets is 600, 100 and 203, respectively.

### 4.2 Features and Evaluation Metrics

The modality features are leveraged in the same way as MEMI and MAG. The evaluation metrics are demonstrated as follows: 1) Mean Absolute Error (MAE) (lower is better); 2) Pearson's correlation (corr); 3) Binary Accuracy (Acc-2); 4) F1-Score (F1); 5)7-class Accuracy (Acc-7). Note that, two distinct manners are introduced to represent Acc-2 and F1: 1) negative/non-negative classification attached with the label [-3, 0) and [0, 3] [Zadeh *et al.*, 2018b] 2) negative/positive classification attached with the label [-3, 0) and (0, 3] [Tsai *et al.*, 2019]. The marker -/- is employed to distinguish the distinct strategies, where the left-side value refer to 1) and the right-side value stands for 2).

### 4.3 Comparisons

We introduce the non-attention and attention based models as the baselines. Non-attention based: RNN-based multistage fusion network (RMFN), Bi-directional LSTM (BC-LSTM), Multi-view LSTM (MV-LSTM), Interaction Canonical Correlation Network (ICCN), Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA), Tensor Fusion Network (TFN), Low-rank Multimodal Fusion (LMF), Deep multimodal fusion (DF). Attention-based: Recurrent Attended Variation Embedding Network (RAVEN), Multimodal Adaptation Gate (MAG), Multimodal Cyclic Translation Network (MCTN), Memory Fusion Network (MFN), Multi-attention Recurrent Network (MARN), Multimodal Factorization Model (MFM), Multimodal Explicit Many2many Interactions(MEMI), Multimodal Transformer (MulT), Self-Supervised Multi-task Multimodal model (Self-MM).

### 4.4 Training Details and Model Complexity

The grid-search is performed over the hyper-parameters to find the model with the best validation task loss. The range of key hyper-parameters are summarized as follows: layer [2, 7], tensor rank [2, 8], residual parameter $\alpha$ [0.1, 0.7]. The complexity of baselines built upon Bert are summarized as follows: MulT ($O(M^2 \times d_i^2 \times T_i)$), TFN ($O(d_y \prod_{i=1}^{M} d_i)$), LMF ($O(d_y \times R_w \times \prod_{i=1}^{M} d_i)$), MISA($O(M \times d_i^2 \times T_i)$), MAG($O(\prod_{i=1}^{M} T_i \times d_i)$), ICCN($O(M \times T_i \times d_i^2)$). Note that, the complexity of our 'MMT' is $O(M \times d_i \times T_i(R_w + R_s))$, where $d_i > R_w(R_s)$. Note that our approach use smaller model rather than larger model, while achieving better performance.

## 5 Experiments Results and Analysis

### 5.1 Performance Comparison with State-of-the-art Models

Firstly, we analyze the performance between our MMT and the state-of-the-art baselines. The bottom rows in Table 1 and Table 2 illustrate the results of our model. Note that $\otimes$ from [Tsai *et al.*, 2019]; $\bigtriangleup$ from [Sun *et al.*, 2020]. As shown in Table 1 (CMU-MOSI), we can find that MMT exceeds the previous best MAG on all metrics. Particularly, our MMT outperforms the previous best MISA on the 'Acc-7' by a margin of 5.9%. It is interesting to observe that, our MMT obtains the much better 'MAE' (lower is better) than the two-way cross-modality attention model MulT (Bert), with a large margin of 20.4%. As shown in Table 2 (POM), MMT exceeds the MulT on the 'Corr' of Con by a margin of 14.3%. This indeed implies the superior expressive capability and efficiency of the proposed multiway multimodality attention mechanism and

| 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 |
| $MCTN^{\otimes}$ | 0.909 | 0.676 | 79.3/- | 79.1/- | 35.6 |
| $MFM^{\otimes}$ | 0.951 | 0.662 | 78.1/- | 78.1/- | 36.2 |
| Bert | 0.739 | 0.782 | 83.5/85.2 | 83.4/85.2 | - |
| MulT | 0.871 | 0.698 | -/83.0 | -/82.8 | 40.0 |
| $TFN(Bert)^{\Delta}$ | 0.901 | 0.698 | -/80.8 | -/80.7 | 34.9 |
| $LMF(Bert)^{\Delta}$ | 0.917 | 0.695 | -/82.5 | -/82.4 | 33.2 |
| MulT (Bert) | 0.861 | 0.711 | 81.5/84.1 | 80.6/83.9 | - |
| 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 | - |
| MMT (ours) (Bert) | **0.657** | **0.83** | **85.8/87.0** | **85.8/87.0** | **48.2** |

Table 1: Performances of baselines and MMT based on BERT in CMU-MOSI benchmark.

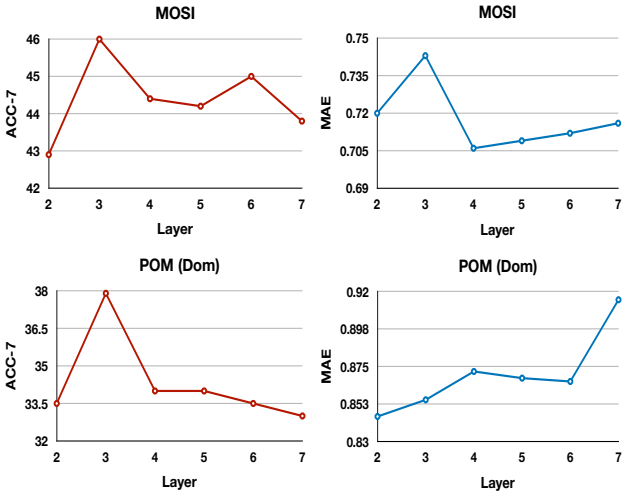


Figure 5: Effect of the layer on CMU-MOSI and POM.

| Task | Ent | Con | Pas | Dom | Viv | Exp |
|---|---|---|---|---|---|---|
| Classes | 7 | 7 | 7 | 7 | 7 | 5 |
| Metric | MAE ↓ | | | | | |
| LSTM | 0.996 | 1.073 | 1.148 | 0.904 | 1.045 | 1.067 |
| BC-LSTM | 0.988 | 1.089 | 1.141 | 0.915 | 1.024 | 1.096 |
| TFN | 1.062 | 1.491 | 1.335 | 1.077 | 1.184 | 1.215 |
| DF | 0.972 | 1.097 | 1.130 | 0.899 | 1.023 | 1.053 |
| MARN | 1.011 | 1.057 | 1.184 | 0.916 | 1.053 | 1.105 |
| MulT | 0.961 | 0.989 | 1.087 | 0.869 | 0.975 | 0.998 |
| MEMI | 0.952 | 0.979 | 1.108 | 0.856 | 0.959 | 0.957 |
| MMT (ours) | **0.911** | **0.941** | **0.987** | **0.845** | **0.944** | **0.934** |
| Metric | Corr ↑ | | | | | |
| LSTM | 0.176 | 0.233 | 0.179 | 0.201 | 0.172 | 0.153 |
| BC-LSTM | 0.083 | 0.200 | 0.219 | 0.318 | 0.241 | 0.177 |
| TFN | 0.265 | 0.159 | 0.158 | 0.067 | 0.232 | 0.149 |
| DF | 0.254 | 0.164 | 0.353 | 0.314 | 0.296 | 0.153 |
| MARN | 0.020 | 0.219 | 0.102 | 0.130 | 0.065 | -0.008 |
| MulT | 0.267 | 0.294 | 0.332 | 0.282 | 0.286 | 0.319 |
| MEMI | 0.275 | 0.432 | 0.327 | 0.395 | 0.392 | 0.365 |
| MMT (ours) | **0.386** | **0.437** | **0.43** | 0.368 | 0.363 | **0.418** |
| Metric | Acc (%) ↑ | | | | | |
| LSTM | 30.5 | 25.1 | 25.1 | 31.5 | 29.6 | 27.6 |
| BC-LSTM | 30.0 | 22.2 | 21.7 | 32.0 | 30.0 | 27.1 |
| TFN | 31.0 | 17.2 | 23.2 | 33.0 | 29.1 | 24.1 |
| DF | 27.6 | 25.1 | 21.7 | 31.5 | 28.6 | 27.1 |
| MARN | 32.5 | 28.6 | 24.6 | 34.5 | 31.0 | 27.6 |
| MulT | 31.5 | 25.1 | 31 | 34 | 35 | 27.6 |
| MEMI | 33.5 | 29.6 | 28.1 | 34.5 | 40.9 | 38.4 |
| MMT (ours) | **34.5** | **33.5** | **34** | **39.9** | 39.4 | 35.5 |

Table 2: Performances of baselines and MMT in POM benchmark.

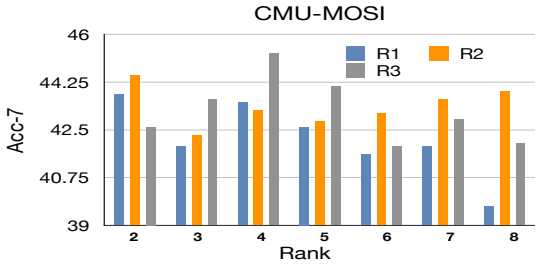


Figure 6: Effect of the tensor ranks on CMU-MOSI.

the hierarchical structure. Compared to the two-way attention model, our presented MMT is capable of exploring the multiway attention space that consists of the many-to-many multimodal interactive path, which naturally benefits the learning model to measure both intermodal and intramodal interactions simultaneously. Essentially, thanks to the low-rank tensor-based strategy, MMT has the ability to effectively perform the multimodal learning task only with a relatively small number of storage and computational complexity.

## 5.2 Effect of the Layer of the Framework

In this part, we attempt to investigate the impact of various layers of the hierarchical framework on the task performance. The layer varies from 2 to 7. As shown in Figure 5, we can observe that MMT reaches the peak value at the layer 3 on the 'Acc-7' on CMU-MOSI. Moreover, the MMT maximizes the performance at the layer 4 on the 'MAE' for the case of CMU-MOSI. As for the Dom of POM benchmark, MMT reaches the highest point at the layer 3 on the 'Acc-7', and the relatively higher performance peak at the layer 2 on the 'MAE'. Intuitively, the relative deeper framework bears the potential to exploit the much more comprehensive and sophisticated sentimental attribution among multiple modality presentations. Subsequently, the deeper cases naturally achieve better performance than the shallow ones. Additionally, the too complex MMT consists of overmuch redundancy message, which may result in the deterioration of the task performance. In conclusion, the experiment results demonstrate that the hierarchical structure indeed is able to provide greater expressive power and efficiency.

## 5.3 Effect of Tensor Ranks of Tensor-ring Network

In this test, we are interested to examine how distinct tensor ranks affect the predictive performance of our tensor-based network. For simplicity, we attempt to analyze the performance of each $r_i$ separately, while the rest of tensor ranks $r_j$ are fixed, where $j \neq i$. For instance, as shown in figure 6, $r_1$ varies from 2 to 8, while $r_2$ and $r_3$ are fixed. In Figure 6, we can observe that the case of $r_1$ and $r_2$ achieve the

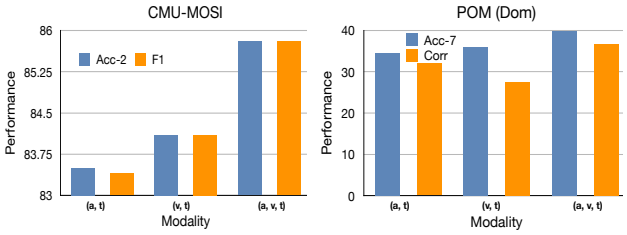


Figure 7: Effect of the number of involved modalities on CMU-MOSI and POM.

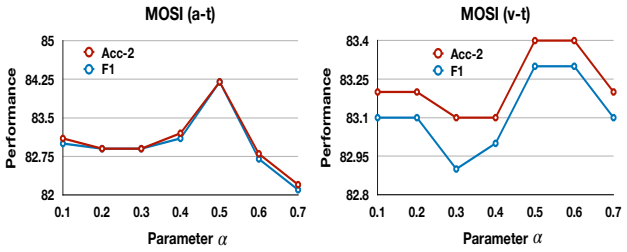


Figure 8: Effect of the residual parameter $\alpha$ on CMU-MOSI.

best performance at rank_2. Similarly, the case of $r_3$ reaches the peak point at the relatively low-rank value rank_4. This implies that the tensor-based fashion indeed gives the learning model sufficient expressive capability to efficiently exploit the task-related representation. Notably, the fairly good results of low-rank cases demonstrate that tensor-based network is able to significantly decrease the computational complexity and storage without the greater deterioration of prediction. The above observations signify the effectiveness and necessity of leveraging the tensor network to deal with the multiway multimodal attention mechanism, when analyzing the case associated with an arbitrary number of modalities.

### 5.4 Effect of the Number of Involved Modalities

Due to the flexibility of our proposed multimodal learning architecture, we can provide a multiway multimodal attention space based on multiple involved modalities. Consequently, in this part, we attend to analyze the distinct framework design associated with the specific number of modalities. As shown in Figure 7, the multimodal case (a, v, t) performs significantly better than the bi-modality cases $\{(a, t), (v, t)\}$. And, the POM depicts similar results. The above performance may indicate that our MMT gives the model strong ability to highlight much more compact consistency and comprehensive complementarity within the multiway multimodal attention space, especially when the model includes much more modality information. Essentially, compared to results of the bi-way cross-modality attention model such as MulT which performed on the multimodal setting (a, v, t) (shown in Table 1), the test bi-modality setting (t, v) of our MMT obtains the comparable 'Acc-2' (84.1) and better 'F1' (84.1) precisions, indicating the great expressive power brought by the proposed multiway attention mechanism. Indeed, the bi-way cross-modality attention mechanism only concerns the intermodal interactions with the limits of structure, while MMT is able to provide both intermodal and intramodal intercorrelations.

### 5.5 Effect of the Residual Parameter $\alpha$

In our work, the residual parameter $\alpha$ is introduced to blend the contribution of modality-aware multiway multimodality message and original modality message. Thus, we attempt to examine how distinct $\alpha$ affects the learning efficiency of the presented model. The $\alpha$ ranges from 0.1 to 0.7. Note that, all the modalities {audio, video, text} share the same $\alpha$. For simplicity, the related testing are performed on the settings $\{$(a, t), (v, t)$\}$. In Figure 8, we can observe that the MMT can reach the good task performance with respect to the tested $\alpha$. In particular, the proposed learning model maximizes the prediction results at the relatively low $\alpha$ 0.5 or 0.6 for all the testing settings. Intuitively, the case associated with the lower $\alpha$ mainly attends to the generated multimodal intercorrelations among the multiple modalities, and largely overlooks the intrinsic sentimental attribution of the modality. Thus, this may result in the lack of sufficient comprehensive multimodal sentiment properties, as well as the deterioration of the task performance. Additionally, the cases associated with the too large $\alpha$ are likely to mainly concentrate on the sentimental information of the original modality presentation, which may introduce too much redundancy to the final multimodal fusion message that may lead to the overfitting issue.

## 6 Conclusion

In this paper, we proposed the tensor-based multiway multimodal transformer for exploring the modality-aware multiway multimodal intercorrelations for each modality in parallel. Indeed, compared to the unidirectional cross-modality attention, MMT is able to accommodate all potential multiway sophisticated interactions from all modalities simultaneously, using the novel multiway multimodal attention. Essentially, the tensor-based network naturally provides us the great flexibility to extend MMT to the case associated with an arbitrary number of modalities. In addition, MMT enjoys great scalability with respect to the number of modalities. In practice, MMT can also serve as the strong baseline to efficiently analyze the relation-extraction tasks, leading to the much more latent and comprehensive relations within the multiway representative space.

## Acknowledgments

This work was supported by National Key R&D Program of China for Intergovernmental International Science and Technology Innovation Cooperation Project (No. 2017YFE0116800), National Natural Science Foundation of China (Grant No.U20B2074, U1909202), JSPS KAKENHI (Grant No.20H04249, 20H04208), and supported by Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province (2020E10010).

## References

[Ain *et al.*, 2017] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and

A Rehman. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, 8(6):424, 2017.

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

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

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

[Nojavanasghari *et al.*, 2016] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 284–288, 2016.

[Park *et al.*, 2014] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 50–57, 2014.

[Pham *et al.*, 2019] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 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, 2019.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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, 2017.

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

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

[Soleymani *et al.*, 2017] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

[Sun *et al.*, 2020] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 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, 2020.

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

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Wang *et al.*, 2019] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.

[Wu *et al.*, 2020] Liangqing Wu, Dong Zhang, Qiyuan Liu, Shoushan Li, and Guodong Zhou. Speaker personality recognition with multimodal explicit many2many interactions. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[Yu *et al.*, 2021] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *arXiv preprint arXiv:2102.04830*, 2021.

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

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

[Zadeh *et al.*, 2018a] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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