Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network

Sijie Mai, Songlong Xing, and Haifeng Hu

Abstract—**The emotion of human is always expressed in a multimodal perspective. Analyzing multimodal human sentiment remains challenging due to the difficulties of the interpretation in inter-modality dynamics. Mainstream multimodal learning architectures tend to design various fusion strategies to learn inter-modality interactions, which barely consider the fact that the language modality is far more important than the acoustic and visual modalities. In contrast, we learn inter-modality dynamics in a different perspective via acoustic- and visual-LSTMs where language features play dominant role. Specifically, inside each LSTM variant, a well-designed gating mechanism is introduced to enhance the language representation via the corresponding auxiliary modality. Furthermore, in the unimodal representation learning stage, instead of using RNNs, we introduce 'channel-aware' temporal convolution network to extract high-level representations for each modality to explore both temporal and channel-wise interdependencies. Extensive experiments demonstrate that our approach achieves very competitive performance compared to the state-of-the-art methods on three widely-used benchmarks for multimodal sentiment analysis and emotion recognition.**

Index Terms—**Acoustic-LSTM, channel-aware temporal convolution, multimodal sentiment analysis, visual -LSTM.**

I. Introduction

Society has witnessed the rapid development of social media and more and more videos that express the speaker's views or emotions are posted online which need to be analyzed carefully and correctly. Intuitively, in these videos, the speaker's opinions are always conveyed in a multimodal perspective [1]. Apart from the spoken language, how language is uttered as well as the accompanied facial gesture should be jointly considered to interpret the exact meaning of the speaker. In other words, language, acoustic and visual are the three essential modalities in human communication. Therefore, multimodal language analysis has become increasingly significant [2]. In this paper, we focus on multimodal sentiment analysis and emotion recognition, which is a common downstream task for multimodal language analysis.

Mainstream multimodal learning methods focus on fusing various representations from different modalities to obtain a joint multimodal representation. Previous works on multimodal sentiment analysis [3], [4] indicate that language modality is far more informative than the visual and acoustic modalities, which should play a dominant role when learning the joint representation. Nevertheless, most of the previous publications treat the three modalities equally which might instead weaken the performance of multimodal system [3], [5]–[8].

Based on this consideration, we innovate to make language modality the dominant role, while the other two modalities serve as auxiliary roles to modify the representation of the language modality. Specifically, we build acoustic- and visual-LSTMs respectively to enhance the language representation successively assisted by acoustic and visual features. Inside each LSTM variant, a well-designed gating mechanism is implemented to dynamically determine whether the auxiliary modalities have discriminative sentiment information and further determine whether the acoustic and visual enhancement should be performed. By the introduction of acoustic- and visual-LSTMs, noisy modalities can be explicitly filtered out and the discriminative sentiment information can be retained to adjust the representation of the language modality.

In addition, temporal convolution network (TCN) [9] has achieved excellent performance and consistently outperformed RNN variants [10]–[12] in a great number of language modeling tasks. Motivated by this observation, unlike previous multimodal sequence learning methods that use recurrent neural networks to learn unimodal representations [1], [5], [7], [13], [14], in this paper we investigate the effectiveness of TCN on modeling unimodal signals. Specifically, we leverage the expressive power of TCN to learn high-level semantic features for each single modality and simultaneously filter out the noisy information by convolutional filters. Moreover, considering that current TCN [9] only applies 1D convolution along the time dimension and pays less attention to explicitly exploring the inner-connection on the channel dimension (i.e., the interaction

Manuscript received February 27, 2020; revised August 27, 2020 and December 31, 2020; accepted March 19, 2021. Date of publication March 26, 2021; date of current version April 20, 2021. This work was supported in part by the National Key Research and Development Program of China under Grants 2018YFB1601101 and 2018YFB1601100, in part by the National Natural Science Foundation of China under Grant 62076262, and in part by the Natural Science Foundation of Guangdong under Grant 2017A030311029. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kai Yu. *(Corresponding author: Haifeng Hu.)*

The authors are with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China (e-mail: maisj@mail2.sysu.edu.cn; xingslong@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TASLP.2021.3068598

within the feature vector at each time step), we introduce 'Channel Interdependency Learning (CIL)' module inside TCN to better model the channel-wise interdependency and thus provide a more powerful representation at each time step.

To sum up, we propose a multimodal learning framework named Temporal Convolutional Multimodal LSTM (TCM-LSTM) to address multimodal human sentiment analysis. The main contributions of this paper are listed below:

- We propose acoustic- and visual-LSTMs to enhance the representation of the spoken language. Inside each LSTM variant, a well-designed gating mechanism is introduced to determine whether the acoustic and visual enhancement should be conducted according to the discriminative information expressed in each modality.
- We introduce a 'channel-aware' temporal convolution network by integrating the 'Channel Interdependency Learning' module into the regular TCN. In this way, our TCN variant is able to explicitly explore both the temporal dependency between time steps and the inner-relation between features within each time step, thereby learning more powerful representations for each modality.
- We conduct extensive experiments to verify that the proposed method achieves state-of-the-art performance across three widely-used datasets for multimodal sentiment analysis. Moreover, the ablation studies and contrast experiments suggest that the proposed components are effective.

II. Related Work

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis has gained a huge attention recently due to the rapid development of social media [1], [2], [15]. In multimodal sentiment analysis, a primary problem is how to fuse the representations from different modalities so as to obtain the multimodal joint representation. Earlier publications mainly include early fusion approaches that extract features from various modalities and conduct modality fusion at input level most by simple concatenation, which show improvement over single modality [8], [16]–[18]. For instance, Poria *et al.* propose Bidirectional Contextual LSTM (BC-LSTM) [8] that attempts to grasp the contextual information from the concatenated multimodal representation. In contrast, late fusion methods firstly infer decision according to each single modality and then weighted average the decisions from all modalities (also called decision-level fusion) [19]–[22]. Nevertheless, these two strategies do not allow the cross-modality interactions to be effectively modeled as elaborated by Zadeh *et al.* [5].

Recently, performing tensor-based fusion to learn inter-modality dynamics has become increasingly popular [23]. Tensor Fusion Network (TFN) [5] adopts outer product to learn the joint representation of three modalities, which is followed by Liu *et al.* [6] and Barezi *et al.* [24] that try to improve efficiency by decomposing weights of the high-dimensional tensor. More recently, Mai *et al.* [3] propose a 'Divide, Conquer and Combine' strategy to conduct local tensor and global fusion, and it is later extended in [14] that uses an elaborately-designed Bidirectional Multi-connected LSTM. Similarly, Hierarchical Polynomial Fusion Network (HPFN) [25] is established to recursively integrate and transmit the local correlations into global correlations by multilinear fusion. Furthermore, some modality translation methods [26] such as Multimodal Cyclic Translation Network (MCTN) [4] aims at learning joint multimodal representation by translating the source modality into the target modality. In addition, a few approaches such as Multimodal Factorization Model (MFM) [27] and Multimodal Baseline (MMB) [28] use factorization methods to learn the multimodal embedding. However, these approaches cannot explicitly outstand the dominant role of the language modality.

More advanced approaches are proposed to learn joint representations of multimodal language at word level [29]–[31]. For example, Recurrent Multistage Fusion Network (RMFN) [1] decomposes multimodal fusion into three stages and performs word-level fusion using LSTM. Moreover, Memory Fusion Network (MFN) [7] implements the delta-memory attention and multi-view gated memory network to fuse memories of LSTMs and explore cross-modal interactions across time. Multi-attention Recurrent Network (MARN) [32] is developed for human communication comprehension to discover cross-modal interactions through time using the multi-attention block. Our baseline, i.e., Recurrent Attended Variation Embedding Network (RAVEN) [13] models multimodal human language by shifting word representations based on facial expressions and vocal patterns, which simply modifies word embedding by integrating the nonverbal shift vector into the original word embedding. In contrast, we implement well-designed acoustic- and visual-LSTM to enhance the word representation successively in a more explicit and powerful way.

In terms of the unimodal representation learning, firstly for the language modality, word2vec [33], GloVe word embeddings [34] and BERT [35] are increasingly popular. For the acoustic modality, many approaches [8], [36], [37] apply openSMILE [38] to extract acoustic features that consist of low-level acoustic descriptors and derivations (LLDs) such as MFCCs and MFSCs [39]. Recently, more researchers [5], [6], [24] tend to adopt COVAREP [40] to extract acoustic features. For visual modality, FACET [1] and 3D-CNN [41] are the main tools to extract features in video. Simple neural networks such as a shallow encoder [5], [24] and LSTMs [8], [11] are used to further process the low-level features to obtain more representative representations, but they are usually not expressive enough. In contrast, based on low-level features, we apply channel-aware temporal convolution network to explicitly learn the high-level channel-wise associations and temporal dependency for each modality.

B. Temporal Convolution Network

Convolutional networks have also been widely used for sequence processing [42]–[44]. Yin *et al.* [45] first conduct a systematic comparison of CNN and RNN on a wide range of representative NLP tasks. Recent publications suggest that temporal

[1] iMotions 2017. https://imotions.com/

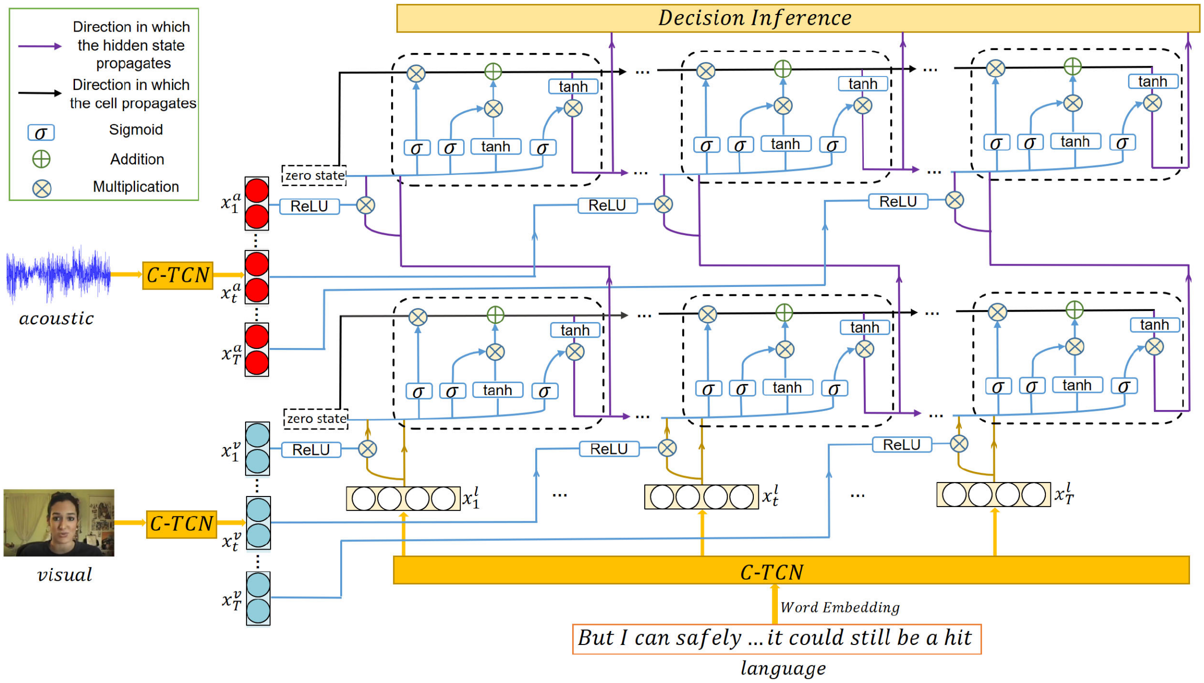

Fig. 1. Schematic Diagram of Our TCM-LSTM.

convolution network (TCN) achieves excellent performance in a variety of NLP tasks, particularly for the ones that demand long-range information propagation [9], [46]. More recently, temporal convolution has been applied on various sequence learning tasks, including video summarization [47], speech enhancement [48], sequential grouping of speaker separation [49], high-dimensional time series forecasting [50], etc. But few of them try to extend TCN by learning the interdependency on the channel (spatial) dimension. In comparison, in this paper we extend TCN to a channel-aware version by introducing the CIL module to better explore the channel-interdependency within each time step.

III. Model Architecture

In this section, we introduce the pipeline of Temporal Convolutional Multimodal LSTM (TCM-LSTM) in detail. As shown in Fig. 1, TCM-LSTM consists of three components: 1) Unimodal Representation Learning module (**URL**) for modeling intra-modality dynamics; 2) Joint Representation Learning module (**JRL**) for learning a joint representation for all modalities; 3) Decision Inference Module (**DIM**) for obtaining the final prediction.

A. Unimodal Representation Learning

URL functions as feature extractor for each single modality. Drawing inspiration from the recent success of utilizing the temporal convolution to extract representative and robust features for sequential data [9], [46], instead of using RNNs to learn the context-dependent information, we leverage the expressive power of the convolutional operation to explore temporal relations across time steps. Moreover, to provide a more powerful way to explore the channel-wise interdependency between features so as to learn better embedding in each time step, we innovate to integrate CIL module into TCN.

There are normally three modalities in multimodal language analysis, whose raw inputs are denoted as $\boldsymbol{u}^a \in \mathbb{R}^{T\times d_a}$ (acoustic), $\boldsymbol{u}^v \in \mathbb{R}^{T\times d_v}$ (visual) and $\boldsymbol{u}^l \in \mathbb{R}^{T\times d_l}$ (language), respectively. Firstly we introduce the regular pipeline of TCN. TCN is basically a series of 1D convolution operated in the time domain which can effectively learn long-range temporal information. Formally, given an input $\boldsymbol{u}^m \in \mathbb{R}^{T\times d_m}$ (where $m \in \{a, v, l\}$), temporal convolution can be expressed as:

$$\boldsymbol{z}_t^m = \boldsymbol{f} \star \boldsymbol{u}_t^m = \sum_{i=0}^{k-1} \boldsymbol{f}_i \boldsymbol{u}_{t-i}^m, 1 \le t \le T \qquad (1)$$

where $\star$ denotes the convolution operation, $\boldsymbol{u}_t^m \in \mathbb{R}^{d_m\times 1}$ is the input feature vector of modality m at time step t, $\boldsymbol{f} \in \mathbb{R}^{k\times d_{out}^m \times d_m}$ is the 1D convolutional kernel with d_m and d_{out}^m being the number of input and output channels respectively, $\boldsymbol{f}_i \in \mathbb{R}^{d_{out}^m \times d_m}$ is the i^{th} dimension of the convolutional kernel, and k is the kernel size. In other words, d_{out}^m is the output dimensionality for each time step. Note that in practice, padding is introduced to retain the same time dimensionality ($\boldsymbol{u}_i^m = \boldsymbol{0}\; if\; i \le 0$) such that $\boldsymbol{z}^m = [\boldsymbol{z}_1^m, \ldots, \boldsymbol{z}_t^m, \ldots, \boldsymbol{z}_T^m]^R \in \mathbb{R}^{T\times d_{out}^m}$, where R denotes the matrix transpose operation. Note that to learn the contextual information from future time steps, our version of TCN is not causal (see Fig. 2), and one can easily extend it to a causal version by adjusting the padding location. In addition, dilated convolution [51], [52] is introduced in [9] to enlarge the receptive field so as to enable very long effective

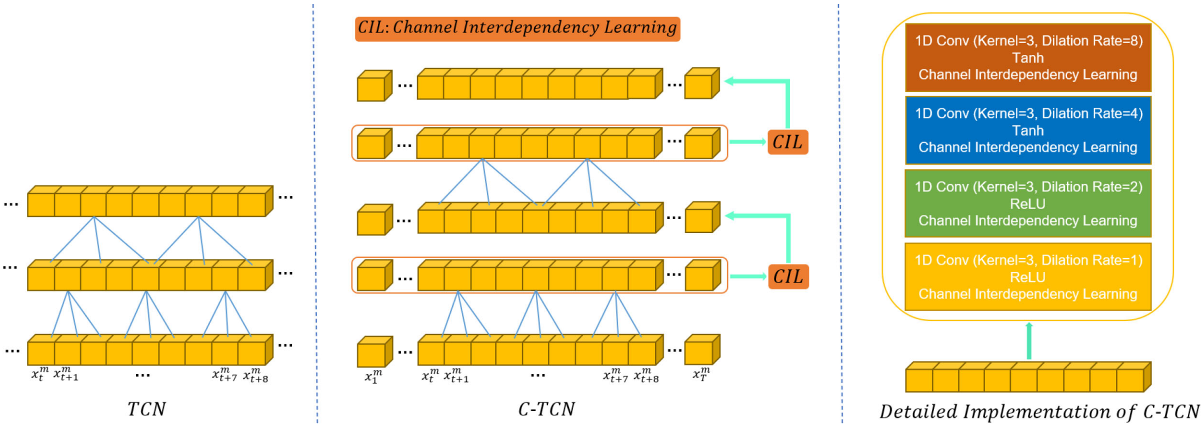
CIL: Channel Interdependency Learning
CIL
CIL
x_t^m x_{t+1}^m ... x_{t+7}^m x_{t+8}^m
TCN
x_1^m x_t^m x_{t+1}^m ... x_{t+7}^m x_{t+8}^m x_T^m
C-TCN
1D Conv (Kernel=3, Dilation Rate=8)
Tanh
Channel Interdependency Learning
1D Conv (Kernel=3, Dilation Rate=4)
Tanh
Channel Interdependency Learning
1D Conv (Kernel=3, Dilation Rate=2)
ReLU
Channel Interdependency Learning
1D Conv (Kernel=3, Dilation Rate=1)
ReLU
Channel Interdependency Learning
Detailed Implementation of C-TCN

Fig. 2. Comparison between TCN and C-TCN. Compared to the regular TCN, our model adds a CIL module after each temporal convolutional layer.

history with reasonable parameters:

$$\boldsymbol{z}_t^m = \boldsymbol{f} \star \boldsymbol{u}_t^m = \sum_{i=0}^{k-1} \boldsymbol{f}_i \boldsymbol{u}_{t-d\cdot i}^m \tag{2}$$

where d is the dilation factor, and $\cdot$ denotes multiplication of scalars.

Nevertheless, TCN mainly focuses on exploring the temporal dependency between time steps, while pays less attention to explicitly exploring the interaction within features at each time step and thus may fail to learn better representation at word (time-step) level. In order to mitigate this problem, motivated by SENet [53], we introduce 'channel interdependency learning (CIL)' operation and construct a 'channel-aware' TCN (C-TCN) (see Fig. 2 for the detailed implementation). In the CIL module, firstly we apply a $ReLU$ activated Fully Connected (FC) layer: $\mathbb{R}^{T\times d_{out}^m} \rightarrow \mathbb{R}^{1\times d_{out}^m}$ on the time dimension so that we can obtain a unified embedding for each channel using features within the channel:

$$\boldsymbol{e}_c = ReLU(\boldsymbol{W}_{fc}\boldsymbol{z}^m) \tag{3}$$

where $\boldsymbol{z}^m \in \mathbb{R}^{T\times d_{out}^m}$, $\boldsymbol{e}_c \in \mathbb{R}^{1\times d_{out}^m}$ and $\boldsymbol{W}_{fc} \in \mathbb{R}^{1\times T}$. Each scalar in the vector $\boldsymbol{e}_c$ has the receptive field across the entire time dimension of the specific channel, which thereby can be regarded as a unified embedding for this channel. Note that by using this learnable method to obtain $\boldsymbol{e}_c$, the C-TCN cannot handle variable-length sequences. Nevertheless, we can still use global average/max pooling on the time dimension to obtain $\boldsymbol{e}_c$ or use padding in the time dimension to ensure that each sequence shares the same length.

Then the unified channel embedding vector $\boldsymbol{e}_c$ is further processed by two FC layers on the channel dimension, where the first ELU [54] activated FC layer is utilized to squeeze the number of channels d_{out}^m by a channel squeeze factor c and the second hard sigmoid activated FC layer is applied to restore the number of channels to the original size, i.e., d_{out}^m. Instead of merely using one FC layer: $\mathbb{R}^{1\times d_{out}^m} \rightarrow \mathbb{R}^{1\times d_{out}^m}$ to compute weights, using two FC layers can not only significantly reduce the number of parameters, but also provide more powerful nonlinearity modeling capacity. The influence of the channel squeeze factor c on model's performance and the number of parameters will be discussed in section IV-G3. In this way, we can obtain the final weight $\boldsymbol{w}_c \in \mathbb{R}^{1\times d_{out}^m}$ for channels, and we repeat the weight vector $\boldsymbol{w}_c : \mathbb{R}^{1\times d_{out}^m} \rightarrow \mathbb{R}^{T\times d_{out}^m}$ to match the dimensions of $\boldsymbol{z}^m$. Finally, we compute the final output using the following equation:

$$\boldsymbol{x}^m = tanh(\boldsymbol{w}_c \odot \boldsymbol{z}^m) \tag{4}$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{w}_c \in \mathbb{R}^{T\times d_{out}^m}$ is the channel weight, $\boldsymbol{z}^m \in \mathbb{R}^{T\times d_{out}^m}$ is the representation for modality m, and $\boldsymbol{x}^m \in \mathbb{R}^{T\times d_{out}^m}$ is the output unimodal representation. The above operation is inspired by the 'channel attention' in SENet [53], but distinction remains: instead of global average pooling, we compute the unified embedding $\boldsymbol{e}_c$ for channels using a learnable FC layer: $\mathbb{R}^{T\times d_{out}^m} \rightarrow \mathbb{R}^{1\times d_{out}^m}$ which is more powerful and expressive. Moreover, 'channel attention' only learns weights for channels, which we think is insufficient to explore interaction within each time step. Therefore, we add the bias term $\boldsymbol{b}$ in (5) which can also be regarded as a kind of residual learning [55]:

$$\begin{aligned} \boldsymbol{b} &= f(f(\boldsymbol{z}^m\boldsymbol{W}_{b1})\boldsymbol{W}_{b2}) \\ \boldsymbol{x}^m &= tanh(\boldsymbol{w}_c \odot \boldsymbol{z}^m + \boldsymbol{b}) \end{aligned} \tag{5}$$

where $\boldsymbol{W}_{b1} \in \mathbb{R}^{d_{out}^m \times \frac{d_{out}^m}{c}}$, $\boldsymbol{W}_{b2} \in \mathbb{R}^{\frac{d_{out}^m}{c} \times d_{out}^m}$, and f is a nonlinear activation function that aims to increase the expressive power. In practice, we empirically choose ELU [54] as it provides the best results. The calculation of the bias term $\boldsymbol{b} \in \mathbb{R}^{T\times d_{out}^m}$ is different from $\boldsymbol{w}_c$ in that we do not apply (3) to squeeze the time dimension so that it can be directly added with $\boldsymbol{w}_c \odot \boldsymbol{x}^m$. In this way, we obtain high-level representation for each time step (i.e., more robust timestep-level representation) by exploring the interdependency within timestep-level feawture vector. The schematic diagram for CIL is shown in Fig. 3 while the complete structure of C-TCN is illustrated in Fig. 2. To retain generalization ability and elegance, the structures of C-TCNs are the same across all modalities, except for the number of output channels (i.e., the dimensionality of the embedding in each time step).

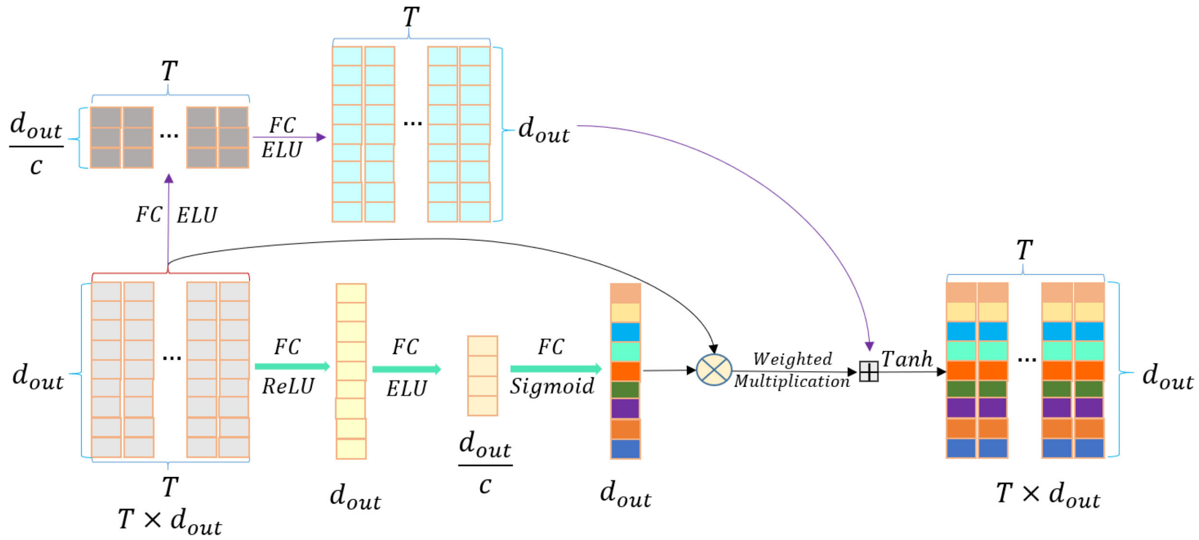

Fig. 3. Schematic Diagram of Channel Interdependency Learning Operation in C-TCN. Note that our version of TCN is non-causal.

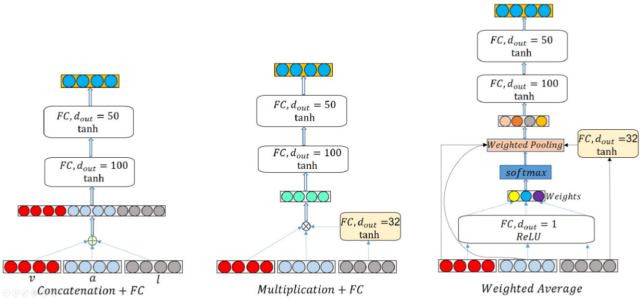

Fig. 4. Schematic Diagram of ‘Concatenation + FC,’ ‘Multiplication + FC’ and ‘Weighted Average’. The figure is modified from [65]. In the ‘Multiplication + FC’ and ‘Weighted Average’ cases, an additional FC layer with output dimensionality 32 is applied for language modality so that it can be directly multiplied or added with the other modalities whose dimensionality for each time step is 32.

B. Joint Representation Learning

In the JRL section, we learn joint multimodal representation by enhancing the language representation via the proposed acoustic- and visual-LSTM.

Since acoustic-LSTM and visual-LSTM share basically the same structure except for the auxiliary modality, we take acoustic-LSTM as an example to illustrate the pipeline. Firstly, we show how we determine whether acoustic enhancement should be conducted at each time step:

$$\boldsymbol{G}_t^a = f(\boldsymbol{W}_1^a \boldsymbol{x}_t^a), \quad 1 \leq t \leq T \tag{6}$$

where f is the ELU [54] activation function, $\boldsymbol{x}_t^a \in \mathbb{R}^{d_{out}^a \times 1}$ is the feature vector at time step t for acoustic modality, $\boldsymbol{W}_1^a \in \mathbb{R}^{h \times d_{out}^a}$ is the learnable parameter matrix. Then, the gating mechanism for auxiliary modality can be expressed in the following equations:

$$\begin{aligned} g_t^a &= ReLU(\sigma(\boldsymbol{W}_2^a \boldsymbol{G}_t^a) - w) \\ \boldsymbol{b}_t^a &= g_t^a \boldsymbol{G}_t^a \end{aligned} \tag{7}$$

where w is the threshold value for acoustic modality which is set to 0.3 in our experiment which provides the best result, $\sigma(\boldsymbol{W}_2^a \boldsymbol{G}_t^a)$ is the gate value of acoustic features at time step t, $\boldsymbol{W}_2^a \in \mathbb{R}^h$ is the learnable parameter, σ is the hard sigmoid activation function to ensure that the value of $\sigma(\boldsymbol{W}_2^a \boldsymbol{G}_t^a)$ is between 0 and 1, and g_t^a is a scalar that represents the weight of acoustic modality at time step t. Obviously, if the gate value $\sigma(\boldsymbol{W}_2^a \boldsymbol{G}_t^a)$ is smaller than the threshold a (in which case acoustic features express little discriminative information), g_t^a will become zero after the $ReLU$ activation function so that the acoustic features will have no influence on the language representation. The introduction of the threshold can reduce the influence of the auxiliary modality and ensure the dominant role of the language modality. In this way, the gating mechanism maintains continuous gradient and simple operation, while at the meantime retains strong discriminative power to filter out the noisy information at each time step. This operation is similar to the sparse attention mechanism introduced in [56], but their objectives are quite different. g_t^a is then multiplied by $\boldsymbol{G}_t^a$ to obtain the final acoustic bias vector $\boldsymbol{b}_t^a$.

After gating mechanism, we modify the routine procedures of LSTM as:

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{f_1} \boldsymbol{x}_t^l + \boldsymbol{W}_{f_2} \boldsymbol{h}_{t-1}^{la} + \boldsymbol{b}_t^a) \tag{8}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{i_1} \boldsymbol{x}_t^l + \boldsymbol{W}_{i_2} \boldsymbol{h}_{t-1}^{la} + \boldsymbol{b}_t^a) \tag{9}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{o_1} \boldsymbol{x}_t^l + \boldsymbol{W}_{o_2} \boldsymbol{h}_{t-1}^{la} + \boldsymbol{b}_t^a) \tag{10}$$

$$\boldsymbol{m}_t = tanh(\boldsymbol{W}_{m_1} \boldsymbol{x}_t^l + \boldsymbol{W}_{m_2} \boldsymbol{h}_{t-1}^{la} + \boldsymbol{b}_t^a) \tag{11}$$

$$\boldsymbol{c}_t^{la} = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}^{la} + \boldsymbol{i}_t \odot \boldsymbol{m}_t$$

$$\boldsymbol{h}_t^{la} = \boldsymbol{o}_t \odot tanh(\boldsymbol{c}_t^{la}) \tag{12}$$

where $\boldsymbol{x}_t^l$ is the language representation at time step t, $\boldsymbol{f}_t$, $\boldsymbol{i}_t$, $\boldsymbol{o}_t$ and $\boldsymbol{m}_t$ refer to forget gate, input gate, output gate, and memory gate, respectively. In the computation of these gates, we add the acoustic bias vector $\boldsymbol{b}_t^a$ to enhance the language representation $\boldsymbol{x}_t^l$ such that the regular LSTM turns into an acoustic-LSTM. Consequently, the hidden state $\boldsymbol{h}_t^{la}$ and the memory state $\boldsymbol{c}_t^{la}$ contain the information of both the language and acoustic modalities. The output of the acoustic-LSTM is a sequence of hidden states $\boldsymbol{h}^{la} = [\boldsymbol{h}_1^{la}, \boldsymbol{h}_2^{la}, \ldots, \boldsymbol{h}_T^{la}]^R \in \mathbb{R}^{T\times h}$ with h being the dimensionality of states. $\boldsymbol{h}^{la}$ is then passed into the visual-LSTM for further fusion:

$$\boldsymbol{h}^{lav} = \text{visual-LSTM}(\boldsymbol{h}^{la}; \boldsymbol{b}_t^v) \tag{13}$$

where $\boldsymbol{b}_t^v$ is the visual bias vector, and $\boldsymbol{h}^{lav}$ is the joint multimodal representation which contains the information of all the three modalities. The visual-LSTM has the same procedure of the acoustic-LSTM except that the auxiliary modality turns into visual modality and the inputs become $\boldsymbol{h}^{la}$ and $\boldsymbol{b}_t^v$. Please refer to Fig. 1 for the detailed structure of acoustic- and visual-LSTM.

C. Decision Inference Module

The Decision Inference Module consists of two LSTM layers followed by two FC layers and dropout layer, which are introduced to explore the temporal dependency and infer the final prediction:

$$\boldsymbol{s} = \mathit{LSTM}(\boldsymbol{h}^{lav}),\ o = \mathit{Dense}(\boldsymbol{s}) \tag{14}$$

where $\boldsymbol{s}$ is the last state of the LSTM network, o is the final output and *Dense* consists of a $tanh$-activated FC layer, a dropout layer and finally a linear FC layer.

IV. Experiments

In this section, we conduct extensive experiments to evaluate TCM-LSTM on multimodal sentiment analysis and emotion recognition on CMU-MOSI [22], IEMOCAP [57], and CMU-MOSEI [58] datasets.

A. Datasets

CMU-MOSI [22] is a popular dataset for multimodal sentiment analysis. The intensity of sentiment in CMU-MOSI ranges from -3 to 3, where -3 denotes the strongest negative sentiment, and +3 the strongest positive. We evaluate our model using various metrics, in agreement with those employed in previous works such as RAVEN [13]. The metrics include 7-class accuracy (Acc7), binary accuracy (Acc2: positive or negative sentiments), F1 score, mean absolute error (MAE) of the score, and the correlation of the model's prediction with human (Corr). To be consistent with prior works, we use 1284 utterances as training set and 686 utterances as testing set.

CMU-MOSEI [58] is the largest multimodal human language analysis dataset so far. The dataset has been segmented at the utterance level. The evaluated metrics for CMU-MOSEI are the same as those for CMU-MOSI. We use 16265 utterances as training set and 4643 utterances as testing set.

IEMOCAP [57] is a multimodal emotion recognition dataset that contains a total of 151 videos from 10 speakers. The videos are segmented into about 10 K utterances. The dataset has the following labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise and other. We take the first four emotions to compare with our baselines. We follow previous works to report the binary classification accuracy and the F1 score of the predictions.

B. Baselines

We compare the performance of TCM-LSTM with the following state-of-the-art multimodal machine learning models:

1) Tensor Fusion Methods: **Tensor Fusion Network** (**TFN**) [5] utilizes 3-fold Cartesian product from unimodal representations to learn unimodal, bimodal and trimodal interactions. **Low-rank Multimodal Fusion** (**LMF**) [6] conducts fusion by decomposing weights of the fusion tensor into a set of low-rank factors to improve efficiency, whose computational complexity scales linearly with the number of modalities. The visual and acoustic features used by TFN and LMF are obtained by performing mean pooling on the original features in the time dimension and thus these features are utterance-level. **Hierarchical Polynomial Fusion Network** (**HPFN**) [25] recursively integrates and transmits the local correlations into global correlations by multilinear fusion. **Hierarchical Feature Fusion Network** (**HFFN**) [3] uses a 'Divide, Conquer and Combine' strategy to conduct local tensor and global fusion, which utilizes sequence-level information using the Attentive Bidirectional Skip-connected LSTM (ABS-LSTM).

2) Word-Level Fusion Methods: **Recurrent Attended Variation Embedding Network** (**RAVEN**) [13], models multimodal human language by shifting word representations based on the facial and vocal features. **Memory Fusion Network** (**MFN**) [7] learns time-dependent cross-modal interactions using Delta-memory Attention Network and Multi-view Gated Memory Network, and it explores modality-specific interactions using systems of LSTMs. **Dynamic Fusion Graph** (**DFG**) [58] extends MFN by using a simple dynamic fusion graph to explore cross-modal interactions. **Recurrent Multistage Fusion Network** (**RMFN**) [1] decomposes fusion into three stages: a HIGHLIGHT stage for highlighting a subset of multimodal sequences, a FUSE stage for performing local fusion of the highlighted features and integrating representations of the previous stage, and a SUMMARIZE stage for drawing the prediction. **Multi-attention Recurrent Network** (**MARN**) [32] is able to discover interactions between modalities through time using the multi-attention blocks and store the interactions in the hybrid memory of a recurrent component. **Multi-view LSTM** (**MV-LSTM**) [59] aims to explore both view-specific and cross-view interactions by partitioning the memory cell and the gates corresponding to multiple modalities.

3) Modality Translation Methods: **Modality Cyclic Translation Network** (**MCTN**) [4] uses Seq2Seq model to learn robust multimodal representations by translating source modality into target modality. MCTN deals with the missing modality problem at testing time which uses the information from source modality to infer prediction. Differently, **Multimodal Transformer** (**MulT**) [26] uses a transformer [60] structure to translate each two modalities.

4) Contextual-Dependent Methods: **Context-aware Interactive Attention** (**CIA**) [61] learns the inter-modality dynamics among the modalities through an auto-encoder mechanism. CIA contains a context-aware attention module to exploit the correspondence among the neighboring utterances, and it uses a bidirectional GRU network to capture sequence-level information. Similarly, **Bidirectional Contextual LSTM** (**BC-LSTM**) [8] also extracts contextual features between utterances. BC-LSTM models the inter-dependencies and relations among the utterances of a video using the bidirectional LSTM, which also utilizes sequence-level information.

5) Factorization Methods: **Multimodal Factorization Model** (**MFM**) [27] factorizes multimodal features into two sets of independent factors, namely multimodal discriminative factors and modality-specific generative factors. MFM is able to learn meaningful multimodal representations and interpret factorized representations to understand the interactions that influence multimodal learning.

TABLE I
COMPARISON BETWEEN TCM-LSTM AND STATE-OF-THE-ART ALGORITHMS ON CMU-MOSI. THE RESULTS LABELED WITH ‡ DENOTES THE SECOND HIGHEST RESULTS, AND THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD. THE RESULTS OF THE METHODS LABELED WITH ⋆ ARE OBTAINED IN OUR OWN EXPERIMENTS

Methods	Acc2	Acc7	F1	MAE	Corr
MV-LSTM [59]	73.9	28.7	74.0	1.040	0.587
GME-LSTM(A) [29]	76.5	-	73.4	0.96	-
DFG [58]	77.7	35.6	77.7	0.96	0.66
BC-LSTM [8]	73.9	28.7	73.9	1.08	0.58
TFN [5]	73.9	32.1	73.4	0.970	0.633
HPFN [25]	77.5	36.7	77.4	0.945	0.672
MARN [32]	77.1	34.7	77.0	0.97	0.63
RMFN [1]	78.4	38.3‡	78.0	0.922	0.681
MFM [27]	78.1	36.2	78.1	0.951	0.662
CIA [61]	79.9	**38.9**	79.5	0.914	0.689‡
MCTN [4]	79.3	35.6	79.1	0.909	0.676
RAVEN [13]	78.0	33.2	76.6	0.915	**0.691**
MFN⋆ [7]	78.7	33.4	78.6	0.969	0.679
MulT⋆ [26]	81.1‡	38.2	81.3‡	0.905‡	0.689‡
HFFN⋆ [3]	80.6	32.8	80.7	0.954	0.656
TCM-LSTM	**81.7**	35.4	**81.8**	**0.903**	0.672

C. Experimental Details

We develop our model on Keras, with tensorflow as backend. The number of output channels of C-TCN, i.e., d_{out} is set to 32, 32 and 100 for visual, acoustic and language modality respectively. The channel squeeze factor c is set to 4. We apply Mean Absolute Error (MAE) as the loss function with Adam [62] of learning rate 0.001 as optimizer. The codes will be made publicly available at[2] if accepted.

In terms of the feature extraction for each modality, to make a fair comparison, we follow the setting of CMU-MultimodalSDK.[3] GloVe word embeddings [34] are used to extract the features of the transcripts in the videos. The Glove word embeddings, represent each word as a 300-dimensional vector, are trained on 840 billion tokens from the common crawl dataset. Facet[4] is used to extract a set of visual features that are composed of facial action units, facial landmarks, head pose, gaze tracking and HOG features [63]. These visual features are extracted from the video utterance at the frequency of 30 Hz to form a sequence of facial gestures over time. COVAREP [40] is utilized for extracting acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity, glottal closure instants, spectral envelope, etc. These acoustic features are extracted from the full audio clip of each utterance at 100 Hz to form a sequence that represents variations in the tone of voice across the utterance.

Since words are the basic units in many natural language processing tasks, P2FA [64] is used for word-level alignment between different modalities such that all modalities are aligned in the time dimension based on the language modality, resulting in the same sequence length across all modalities. The duration of a time step depends on the interval duration of the corresponding word in the utterance, and the sequence length depends on the number of words in the utterance. To ensure that all the utterances in the same dataset have the shared sequence length, padding and cutting operations are introduced, which are responsible for padding and clipping the utterances to the desired length, respectively. The sequence length is set to 50 in CMU-MOSI and CMU-MOSEI datasets, and 20 in IEMOCAP dataset. The extracted features are then fed into the URL module for further processing.

D. Compare With State-of-The-Art Approaches

Results on CMU-MOSI dataset: As we can infer from Table I, our TCM-LSTM shows improvement over typical approaches on the binary accuracy, MAE and F1 score. Compared with the famous BC-LSTM [8] and MFN [7], TCM-LSTM achieves improvement on binary accuracy and F1 score by about 3% and 8% respectively, which demonstrates the superiority of our method. Compared to our baseline RAVEN [13] whose multimodal embedding learning strategy is closest to ours, our TCM-LSTM outperforms it by over 3.5% on binary accuracy, 2% on 7-class accuracy and 5% on F1 score, while is weaker than it in terms of correlation coefficient. Nevertheless, to synthesize all the evaluation metrics, our method still performs better than RAVEN. We argue that it is partly because RAVEN simply modifies word representation by integrating the nonverbal shift vector into the original word embedding. In contrast, we implement elaborately-designed acoustic- and visual-LSTMs respectively to enhance word representation successively in a more explicit and powerful way. Also, instead of LSTMs, we implement channel-aware TCN to extract high-level representation for each modality.

Results on CMU-MOSEI dataset: To evaluate TCM-LSTM's ability in dealing with more challenging multimodal learning tasks, we further report TCM-LSTM's performance on the largest multimodal language analysis dataset CMU-MOSEI and the results are presented in Table II. From the results we can infer that the proposed TCM-LSTM achieves best performance on 7-way classification accuracy and MAE among all the approaches. While in 2-way accuracy and F1 score, our method

[2] https://github.com/TmacMai/multimodal-fusion

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

[4] iMotions 2017. https://imotions.com/

TABLE II
COMPARISON BETWEEN TCM-LSTM AND STATE-OF-THE-ART ALGORITHMS ON CMU-MOSEI. THE RESULTS OF THE METHODS LABELED WITH ⋆ ARE OBTAINED IN OUR OWN EXPERIMENTS. THE PRE-EXTRACTED FEATURES ARE THE SAME ACROSS ALL MODELS.

Methods	Acc2	Acc7	F1	MAE	Corr
DFG [58]	76.9	45.0	77.0	0.71	0.54
MCTN [4]	79.8	49.6	80.6	0.609	0.670
CIA [61]	80.4	50.1‡	78.2	0.680	0.590
RAVEN [13]	79.1	50.0	79.5	0.614	0.662
MulT⋆ [26]	81.1	49.6	81.3	0.616	0.673
MFN⋆ [7]	80.4	49.1	80.3	0.608	**0.679**
HFFN⋆ [3]	**81.5**	49.9	**81.7**	0.607‡	0.675‡
TCM-LSTM	81.4‡	**50.6**	81.6‡	**0.606**	0.673

ranks second and is slightly inferior to HFFN [3] by 0.1 point. To synthesize these results, our TCM-LSTM achieves comparable performance to the state-of-the-art methods on CMU-MOSEI dataset, demonstrating the robustness of our model in terms of handling more challenging tasks. Compared to our baseline RAVEN [13], TCM-LSTM outperforms it on all the evaluation metrics by a significant margin, which further proves the effectiveness of our method.

Results on IEMOCAP dataset: As presented in Table III, our method demonstrates improvement on Angry emotion on both accuracy on F1 score, and it obtains the second highest accuracy and F1 score on Neutral emotion. Moreover, TCM-LSTM achieves the best F1 score and the second highest accuracy on Sad emotion. On Happy emotion, TCM-LSTM obtains the second highest F1 score. These results suggest that our model reaches competitive performance compared to the state-of-the-art models on IEMOCAP dataset.

To synthesize the results of the three datasets, TCM-LSTM consistently reaches excellent performance across multiple datasets, demonstrating the effectiveness and robustness of our model.

E. Model Complexity Analysis

We use the number of trainable parameters as the metric for the model complexity. As presented in Table IV, our model has 332 331 trainable parameters, which is approximately 41.10% and 21.45% of the number of parameters of MFN and MulT, respectively. Nevertheless, compared to HFFN which aims to improve the efficiency of tensor fusion, our model is still clearly more complex in terms of the model complexity. The results suggest that our TCM-LSTM has moderate model complexity and it has much fewer parameters than the current state-of-the-art method MulT, which demonstrates that the improvement of TCM-LSTM over the other models does not simply result from the increase in the number of trainable parameters.

F. Ablation Studies

To investigate the contributions of C-TCN, visual-LSTM and acoustic-LSTM, ablation studies are performed. In each contrast experiment, one certain component is removed or replaced by the other components for comparison. Firstly for C-TCN, we can infer from the 'TCN' case in Table V that the removal of CIL module brings over 1% drop in both the binary accuracy and F1 score. Moreover, the performance on MAE and Corr is also clearly inferior to the model with CIL module, indicating that learning channel-wise interdependency in temporal convolution is significant.

Furthermore, when we remove the acoustic- and visual-LSTMs such that only language information is available, the performance drops dramatically by about 2% in binary accuracy and 3% in 7-class accuracy. Moreover, in the cases of 'Acoustic-LSTM' or 'Visual-LSTM' where only two modalities are utilized (language and acoustic or language and visual), the performances are much better than the case where only the language modality is available, but still are clearly weaker than the case that all the modalities are included. The experiments demonstrate the effectiveness of acoustic- and visual-LSTM in enhancing the language representation. The results also indicate that the bimodal and trimodal systems significantly outperform unimodal system and highlight the necessity of analyzing human language in a multimodal perspective.

In the case of 'Concat-LSTM,' we replace acoustic- and visual-LSTMs with the Concat-LSTM and evaluate the contribution of the gating mechanism in our LSTM variants. Concat-LSTM means simply concatenating the input with the representation of the corresponding auxiliary modality, and then send the new input to a plain LSTM. As we can infer from Table V, Concat-LSTM performs only slightly better than the case where only language information is available ('Remove'), and is clearly weaker than the cases where visual- and acoustic-LSTM are included, even though it has more parameters. It proves our initial motivation that treating all the modalities equally might instead harm the performance of multimodal system and highlights the necessity of the gating mechanism in our LSTM variants.

In addition, we also compare our C-TCN with the TCN version whose number of parameters is approximately the same as C-TCN to estimate whether the improvement of C-TCN results from the increase in the number of parameters. In the case of Deeper TCN (see Table V), we increase the number of layers in TCN for each modality to 5 and 6 respectively where the dilate rate of the 5^{th} and 6^{th} layers is set to 16, and the channel fusion module is removed. From Table V, we can infer that increasing the layers of TCN to 5 layers does help to slightly improve the overall performance, given that the performance on the MAE and Corr metrics is improved evidently. However, the improvement is small compared to our C-TCN, which proves that the superior performance of the C-TCN is not entirely due to the increase in the number of parameters and highlights the effectiveness of our channel fusion module. Moreover, the 5-layer TCN outperforms the 6-layer TCN, suggesting that simply increasing the layers of TCN does not necessarily lead to better results.

Note that in the proposed model, we employ a learnable approach to obtain the unified embedding for channels, as in (3) with a learnable parameter $\mathbf{W}_{fc}$. Due to the fixed size of $\mathbf{W}_{fc}$, this approach deals with the test sequence of greater length than the pre-defined length T by cropping. In other words, only the first T time steps are considered. One intuitive non-learnable way to prevent cropping during testing is by pooling over the time dimension to obtain the unified embedding for channels. We

TABLE III
COMPARISON BETWEEN TCM-LSTM AND OTHER APPROACHES ON IEMOCAP DATASET. THE RESULTS OF THE METHODS LABELED WITH $\star$ ARE OBTAINED IN OUR OWN EXPERIMENTS

	Happy		Sad		Angry		Neutral	
Models	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MV-LSTM [59]	85.9	81.5	81.1	78.8	82.5	82.4	65.2	64.9
BC-LSTM [8]	84.9	81.7	83.2	81.7	83.5	84.2	67.5	64.1
MARN [32]	86.7	83.6	82.0	81.2	84.6	84.2	66.8	65.9
MFN [7]	86.5	84.0	83.5	82.1	85.0	83.7	69.6	69.2
DFG [58]	86.8	84.2	83.8	83.0	85.8	85.5	69.4	68.9
RAVEN [13]	87.3‡	**85.8**	83.4	83.1	87.3	86.7	69.7	69.3
LMF$\star$ [6]	86.9	82.3	**85.4**	84.7	87.1	86.8	**71.6**	**71.4**
MulT$\star$ [26]	**87.4**	84.1	84.2	83.1	88.0‡	87.5‡	69.9	68.4
HFFN$\star$ [3]	86.8	82.1	84.4‡	84.5	86.6	85.8	69.6	69.3
TCM-LSTM	87.2	84.8‡	84.4‡	**84.9**	**89.0**	**88.6**	71.3‡	71.2‡

TABLE IV
THE COMPARISON OF MODEL COMPLEXITY ON CMU-MOSEI. FOR HFFN [3], SINCE IT IS NOT AN END2END METHOD, WE COMPARE THE COMPLEXITY OF ITS MULTIMODAL FUSION PART AS WELL AS THE ENTIRE UNIMODAL LEARNING PART PLUS MULTIMODAL FUSION PART

Methods	Number of Parameters
MulT [26]	1,549,321
MFN [7]	808,421
HFFN (multimodal) [3]	21,980
HFFN (multimodal+unimodal) [3]	261,930
TCM-LSTM	332,331

TABLE V
THE IMPORTANCE OF DIFFERENT COMPONENTS ON CMU-MOSEI. 'TCN' MEANS REPLACING C-TCN WITH REGULAR TCN (REMOVING THE CIL MODULE), 'REMOVE' MEANS REMOVING THE VISUAL- AND ACOUSTIC-LSTMS WHERE ONLY LANGUAGE INFORMATION IS AVAILABLE. 'VISUAL-LSTM' MEANS REMOVING THE ACOUSTIC-LSTM AND ONLY RETAINING THE VISUAL-LSTM SUCH THAT THE INPUT OF VISUAL-LSTM BECOMES LANGUAGE FEATURE INSTEAD OF THE OUTPUT OF ACOUSTIC-LSTM, AND 'ACOUSTIC-LSTM' VICE VERSA. IN ADDITION, 'CONCAT-LSTM' MEANS SIMPLY CONCATENATING THE INPUT WITH AUDIO OR VISUAL FEATURES AS THE NEW INPUT TO ONE PLAIN LSTM AND REPLACING THE ACOUSTIC- AND VISUAL-LSTM WITH IT

Methods	Parameters	Acc2	Acc7	F1	MAE	Corr
TCN	283,539	80.2	50.6	80.5	0.618	0.661
Deeper TCN (5)	319,847	80.3	50.5	80.3	0.612	0.669
Deeper TCN (6)	356,155	80.2	49.9	80.5	0.614	0.668
TCN (mean pooling)	331,731	80.5	**51.1**	80.8	0.611	0.669
TCN (max pooling)	331,731	81.3	50.6	**81.7**	0.609	0.666
Remove	300,411	79.6	47.6	80.2	0.631	0.650
Visual-LSTM	306,435	80.1	49.9	80.6	0.615	0.668
Acoustic-LSTM	310,179	80.6	48.2	80.9	0.629	0.653
Concat-LSTM	338,411	79.9	49.7	80.1	0.618	0.667
TCM-LSTM	332,331	**81.4**	50.6	81.6	**0.606**	**0.673**

experiment with both mean/max pooling and present the results in Table V (see the case of 'TCN (mean pooling)' and 'TCN (max pooling)' in Table V). It can be seen that mean/max pooling both achieve high performance, faring the best in terms of Acc7 and F1 score, respectively. However, our proposed approach still outperforms them on the remaining three metrics, showing the advantage of the learnable channel fusion approach over the non-learnable ones.

TABLE VI
DISCUSSION OF DIFFERENT FUSION STRATEGIES ON CMU-MOSEI DATASET. PLEASE REFER TO THE CORRESPONDING PAPERS [3], [5], [6] FOR THE DIAGRAMS OF TENSOR FUSION, ABS-LSTM AND LOW-RANK MODALITY FUSION RESPECTIVELY. THE OUTPUT DIMENSIONALITY OF ABS-LSTM IS SET TO 16. IN THE 'CONCATENATION + LSTM' APPROACH, WE DIRECTLY CONCATENATE THE REPRESENTATIONS OF ALL THE THREE MODALITIES AND SEND IT INTO A MULTI-LAYER LSTM NETWORK WHOSE OUTPUT DIMENSIONALITY IS SET TO 16. THE DIAGRAMS OF THE OTHER METHODS FOR COMPARISON ARE GIVEN IN FIG. 4. WE ALSO COMPARE THE PROPOSED MODEL WITH VARYING FIXED THRESHOLD w. 'LEARNABLE $\sigma(w^m) = 0.3$' MEANS SETTING THE THRESHOLD AS A LEARNABLE PARAMETER INITIALIZED TO BE 0.3

Methods	Acc2	Acc7	MAE
Concatenation + FC	79.8	48.9	0.624
Multiplication + FC	66.4	41.6	0.804
Weighted Average	80.4	49.3	0.622
Tensor Fusion	80.7	49.3	0.622
Low-rank Modality Fusion	78.0	47.5	0.673
Concatenation + LSTM	80.0	49.8	0.610
Concatenation + ABS-LSTM	80.6	49.9	0.615
Acoustic- and Visual-LSTM ($w = 0.3$)	81.4	50.6	**0.606**
Acoustic- and Visual-LSTM (learnable $\sigma(w^m)$)	80.9	50.3	0.609
Acoustic- and Visual-LSTM (learnable $\sigma(w^m)$= 0.3)	81.2	50.2	0.613
Acoustic- and Visual-LSTM ($w = 0.1$)	80.5	**50.9**	0.613
Acoustic- and Visual-LSTM ($w = 0.2$)	**81.5**	48.9	0.609
Acoustic- and Visual-LSTM ($w = 0.4$)	80.1	49.9	0.613
Acoustic- and Visual-LSTM ($w = 0.5$)	80.4	50.6	0.612
Acoustic- and Visual-LSTM ($w = 0.6$)	79.9	49.1	0.624
Acoustic- and Visual-LSTM ($w = 0.7$)	79.6	49.0	0.625

G. Analysis

1) Discussion of Fusion Strategies:

a) Comparison With Other Fusion Methods: In order to verify that our fusion strategy is indeed effective, we conduct a contrast experiment to compare with other fusion strategies. We can infer from Table VI that our fusion method brings significant improvement on performance compared to other methods, demonstrating its effectiveness and superiority. Specifically, our LSTM variants outperform the baseline 'Concatenation + LSTM' by a significant margin. Furthermore, integrating intra- and inter-attention inside LSTM, the 'Concatenation + ABS-LSTM' strategy [3] is able to achieve very competitive results and significantly outperforms 'Concatenation + LSTM'. Nevertheless, our method still outperforms it on all the evaluation metrics, even through we only add a bias term inside LSTM. We argue that this is because our visual- and acoustic-LSTM explicitly highlights the dominant role of language modality which is not considered in both 'Concatenation + ABS-LSTM' and 'Concatenation + LSTM'.

'Tensor Fusion' [5] is able to reach impressive results, but it requires far more computational resources for training. In contrast, though efficient, the 'Low-rank Modality Fusion' [6] achieves relatively bad performance. Interestingly, the simplest fusion approach 'Concatenation + FC' outperforms advanced fusion strategies 'Low-rank Modality Fusion,' which indicates that a simple fusion method is not necessarily bad. In addition, the 'Weighted Average' fusion method performs better than 'Multiplication + FC' and 'Concatenation + FC,' mainly for the reason that it considers the importance of modalities. From above analysis, we can draw the conclusion that the concrete fusion strategy of our TCM-LSTM is a crucial factor that leads to the marked improvement of performance.

In general, the baselines can be categorized by their motivations. 'Concatenation + FC,' 'Multiplication + FC' and 'Weighted Average' fall into the first category, which is the most intuitive fusion method relying on simple baseline fusion methods followed by fully connected layers. The second category subsumes 'Tensor Fusion' and 'Low-rank Modality Fusion,' which are both tensor-based methods. Tensor fusion is one widely adopted approach among the mainstream fusion methods, offering one type of strong competing baselines to our proposed model. This category does not acknowledge the various importance of different modalities. The third category includes 'Concatenation+LSTM' and 'Concatenation+ABS-LSTM,' which also employs LSTM to explore contextual dependency but without discerning the modality importance. This category is the most direct comparison to our proposed fusion method to verify our motivation of emphasizing the language modality. Overall, the third category performs best among the other two, showing the effectiveness of LSTM. The remaining two categories achieve the similar performance. However, all of the three categories do not yield comparable performance to our Acoustic- and Visual-LSTM. This shows the importance of our motivation of discerning the various modality importance, and the effectiveness of the devised intelligent gate mechanisms. We also conduct extensive analysis on the representative test samples to reveal the effectiveness of Acoustic- and Visual-LSTM in section IV-G4 and IV-G5.

b) Analysis on the Threshold Values: Additionally, we investigate the situation that the threshold value w in our gating mechanism is learnable instead of empirically chosen (see (7)). The equation for obtaining g_t^m thus becomes $g_t^m = ReLU(\sigma(\boldsymbol{W}_2^m \boldsymbol{G}_t^m) - \sigma(w^m)),\ m \in \{a, v\}$, where w^m is the learnable threshold for modality m and the hard sigmoid activation function σ ensures that $\sigma(w^m)$ falls in the range [0, 1]. As presented in Table VI, the learnable threshold version of TCM-LSTM achieves very competitive results compared to other fusion baselines, but still is slightly weaker than the original TCM-LSTM. Note that the final value of $\sigma(w^m)$ becomes 0 and 0.42 for acoustic and visual modality respectively. Moreover, we implement a version where $\sigma(w^m)$ is initialized as 0.3 for both modalities and learned via gradient descent, and this version provides a higher binary accuracy compared to the randomly initialized one. The final value of $\sigma(w^m)$ is 0.32 and 0.24 for acoustic and visual modality, respectively, which slightly drifts from 0.3. Furthermore, we also provide the results for the other fixed thresholds ranging from 0.1 to 0.7. The results suggest that a large threshold ($w = 0.6$ or 0.7) leads to unfavorable performance compared to the medium values of threshold. This is reasonable because a large threshold means that a great amount of information for the acoustic and visual modalities is lost.

2) Discussion of Unimodal Representation Learning: In this section, we conduct a contrast experiment to verify that our unimodal representation learning architecture, i.e., C-TCN, outperforms RNN variants. The baselines for comparison include the widely-used GRU [12] and LSTM [11] networks, as well as the more advanced RNN variants that aim to learn longer temporal dependency such as Recurrent Highway Network (RHN) [66] and Highway LSTM [67]. As we can infer from Table VII, when we replace C-TCN with RNN variants, the performance of the model is still very impressive, demonstrating the effectiveness of RNN variants on modeling sequences. Nevertheless, the model's performance still decreases compared to our TCM-LSTM. To be more specific, combining the results of all the five evaluation metrics, Highway LSTM reaches the best performance among all the RNN variants (except for the bidirectional LSTM which we will discuss later) and even outperforms the simple TCN (see Table V for the result of simple TCN). But it is weaker than our C-TCN in terms of all the metrics. In addition, training RNN variants needs more computational resources, which is a huge disadvantage of RNN variants against TCN variants. For instance, a total number of 691, 827 parameters are required for training Highway LSTM in the architecture of TCM-LSTM, while training C-TCN only requires 332, 331 parameters.

Moreover, we also investigate the effectiveness of bidirectional LSTM on modeling unimodal sequences, which utilizes the sequence-level information similar to our method. The results suggest that the bidirectional LSTM's performance is very competitive. Interestingly, the results of bidirectional LSTM and C-TCN are very close in terms of binary accuracy, F1 score, MAE, and Corr. While on 7-class classification accuracy, our C-TCN outperforms the bidirectional LSTM by 1.1 points. Therefore, to synthesize the results from the five metrics, we argue that C-TCN slightly outperforms bidirectional LSTM. Moreover, bidirectional LSTM requires far more parameters to be trained compared to our C-TCN. We also implement a light-weight version of bidirectional LSTM which has a similar number of parameters to C-TCN. It can be seen that it achieves satisfactory performance. Nonetheless, it still compares unfavorably to the original bidirectional LSTM, which shows that LSTM indeed requires a larger number of learnable parameters to achieve comparable performance to C-TCN.

3) Discussion of Channel Squeeze Factor c: To investigate how the number of parameters and the performance of the C-TCN changes with channel squeeze factor c, a contrast experiment is performed. As shown in Fig. 5, the number of parameters decreases with the increase of c, but the extent of the decrease in

TABLE VII
DISCUSSION OF UNIMODAL REPRESENTATION LEARNING ON CMU-MOSEI. WE REPLACE EACH C-TCN LAYER WITH THE CORRESPONDING RNN/TCN VARIANT AND MAINTAIN THE SAME OUTPUT DIMENSIONALITY IN THIS EXPERIMENT. FOR THE BIDIRECTIONAL LSTM, THE OUTPUT DIMENSIONALITY IS TWICE AS LARGE AS THAT OF THE C-TCN. FOR FAIR COMPARISON, WE IMPLEMENT A LIGHT-WEIGHT VERSION OF BIDIRECTIONAL LSTM WHOSE NUMBER OF PARAMETERS IS APPROXIMATELY THE SAME AS OUR C-TCN (BY REDUCING THE OUTPUT DIMENSIONALITY OF THE BIDIRECTIONAL LSTM). THE HIGHWAY-LSTM [67] FOR COMPARISON ADDS HIGHWAY NETWORK [68] IN MEMORY CELL. THE NUMBER OF HIGHWAY LAYERS IN RHN [66] IS SET TO 2 AS IT PROVIDES THE BEST PERFORMANCE

Methods	Acc2	Acc7	F1 score	MAE	Corr	Number of Parameters
GRUs [12]	80.6	49.7	81.0	0.611	0.675	427,459
LSTMs [11]	81.0	49.4	81.1	0.612	0.675	545,283
Recurrent Highway Network [66]	80.9	50.2	81.2	0.610	0.669	502,403
Highway LSTM [67]	81.0	50.3	81.3	0.615	0.658	691,827
Bidirectional LSTM [11]	**81.5**	49.5	81.5	**0.606**	0.675	1,314,627
Bidirectional LSTM (light-weight) [11]	81.1	49.3	81.5	0.609	**0.684**	353,923
C-TCN	81.4	**50.6**	**81.6**	**0.606**	0.673	**332,331**

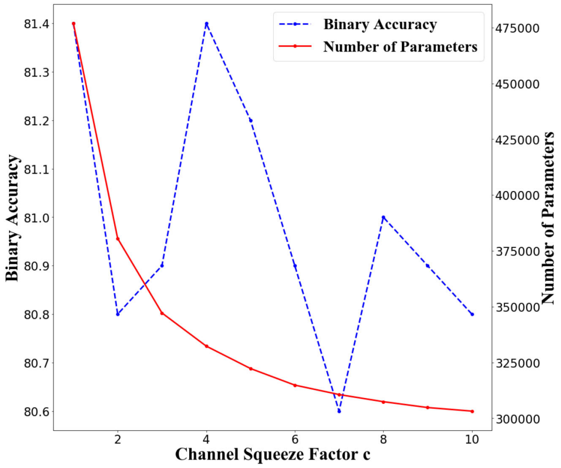

Fig. 5. Influence of Channel Squeeze Factor c.

the number of parameters gradually decline. As for the binary accuracy, actually the model performs robustly as c changes from 1 to 10, with no more than 1% fluctuation. These results suggest that our model has good parametric stability with respect to c. Specifically, when c is set to 1 and 4, the model reaches the best performance. For the sake of model complexity, we choose 4 as the final value of the channel squeeze factor c.

4) Visualization for the Gate Values of Visual and Acoustic Features: In this section, we provide an analysis on the gate values of the visual/acoustic features (see (7)) to investigate when visual/acoustic information takes effect or otherwise. An utterance is shown in Fig. 6, whose sentiment score is -2.2. For visual features, obviously the speaker uses a smiling face during the middle of the utterance which is contradictory to the sentiment score as well as the sentiment of language modality. Our gating mechanism successfully predicts the gate values of this part of the visual modality as less than the threshold 0.3, such that our model is not negatively influenced by the misleading visual modality to predict this utterance as positive. For acoustic features, the speaker generally uses a tone that expresses slightly negative sentiment, and the gate values of acoustic modality are consistently higher than the threshold 0.3. With the help of the gating mechanism, our model predicts the sentiment score of this utterance as -1.53, which is rather close to the true sentiment score. This example suggests that our gating mechanism can filter out the visual/acoustic information that is harmful to the multimodal system and assign a higher weight to the useful information.

5) Case Study for Multimodal Language: One positive example of our multimodal system is shown in Instance 1 of Fig. 7. When the visual- and acoustic-LSTM are removed, our system classifies this utterance as weak positive possibly because of the word ‘welcome’ in the transcript. But in fact, the spoken sentence does not convey much sentiment information clearly. However, as can be seen from the figure, the speaker shows a disgusted face and uses a mocking voice which expresses negative sentiment. By introducing the visual- and acoustic-LSTM, our model finally infers it as weak negative which is consistent with the true label. It demonstrates the effectiveness of multimodal system in the circumstance where language modality expresses little emotional information.

A negative sample is shown in Instance 2 of the Fig. 7. This instance’s sentiment score is -0.6 (negative), while the model with visual- and acoustic-LSTM predicts it as 2.04 (positive). In this instance, the speaker uses an excited and buoyant voice. Meanwhile, from the figure one can tell that the speaker has a very positive expression. However, from the transcript it can be inferred that his opinion towards the movie is somewhat negative. Nevertheless, our system misclassifies it as positive possibly based on the visual and acoustic cues, which indicates that our system still does not perfectly cope with the situation where modalities deliver contradictory information. How to cope with the situations where different modalities express contradictory information remains a challenging task.

Instance 2 inspires us to investigate how the model performs when the language modality expresses contradictory emotion to visual/acoustic information. Actually, we analyze the first 100 testing utterances in the CMU-MOSI dataset, and find that there are five utterances in which the acoustic/visual modalities convey contradictory information to language modality, suggesting that the situation where different modalities express contradictory sentiment is rare. In these five utterances, our TCM-LSTM misclassifies one utterance in the binary classification task. In contrast, the MulT [26] and MFN [7] misclassifies two and three utterances, respectively. The results to some extent demonstrate that our TCM-LSTM handles contradictory utterances better than the current SOTA methods.

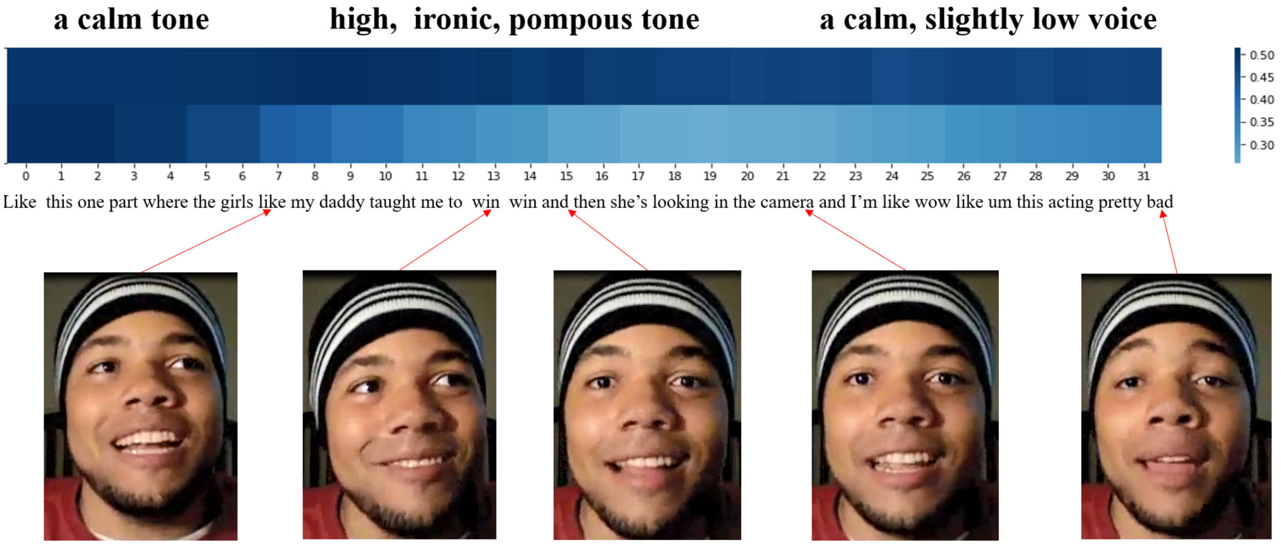

Fig. 6. Visualization for the Gate Values of Visual and Acoustic Features. The first and the second row denotes the gate values for acoustic and visual features, respectively.

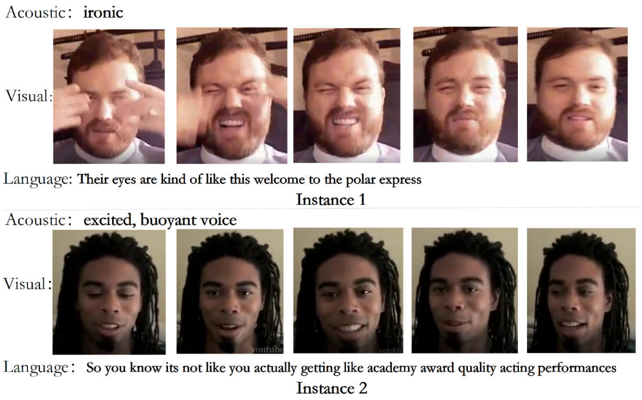
Acoustic: ironic
Visual:
Language: Their eyes are kind of like this welcome to the polar express
Instance 1
Acoustic: excited, buoyant voice
Visual:
Language: So you know its not like you actually getting like academy award quality acting performances
Instance 2

Fig. 7. Positive and Negative Instances for Multimodal Language. Here, Instance 1 is correctly classified into negative sentiment after applying multimodal learning system, while Instance 2 is misclassified into positive sentiment.

V. Conclusion

In this paper, we learn multimodal embedding that enhances language representation in a different perspective via acoustic- and visual-LSTM where language features play dominant role. In addition, we innovate to design a 'channel-aware' temporal convolution operation to extract high-level representations for each modality by exploring both temporal and channel-wise interdependency. Extensive experiments demonstrate that our model reaches very competitive results across multiple benchmark datasets.

References

[1] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 150–161.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern, Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[3] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. 57th Ann. Meeting Assoc. Comput. Linguistics*, Jul. 2019, pp. 481–492.

[4] H. Pham, P. P. Liang, T. Manzini, L. P. Morency, and P. Barnabǎs, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.

[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1114–1125.

[6] Z. Liu, Y. Shen, P. P. Liang, A. Zadeh, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[7] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[8] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L. P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 873–883.

[9] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.

[10] M. W. Goudreau, C. L. Giles, S. T. Chakradhar, and. D. Chen, "First-order versus second-order single-layer recurrent neural networks," *IEEE Trans. Neural Netw.*, vol. 5, no. 3, pp. 511–513, May 1994.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural, Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[13] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 7216–7223.

[14] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 122–137, Jan. 2020.

[15] S. Mai, S. Xing, J. He, Y. Zeng, and H. Hu, "Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion," 2020, *arXiv:2011.13572*.

[16] M. Wollmer *et al.*, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May/Jun. 2013.

[17] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of SVM trees for multimodal emotion recognition," in *Proc. Signal Inf. Process. Assoc. Summit Conf.*, 2012, pp. 1–4.

[18] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining*, 2016, pp. 439–448.

[19] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affective Comput.*, vol. 2, no. 1, pp. 10–21, Jan./Jun. 2011.

[20] B. Nojavanasghari, D. Gopinath, J. Koushik, and L. P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 284–288.

[21] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," in *Proc. Assoc. Comput. Linguistics Short Paper*, 2018, pp. 606–611.

[22] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016.

[23] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proc. Assoc. Comput. Linguistics*. Florence, Italy: Assoc. Comput. Linguistics, Jul. 2019, pp. 1569–1576.

[24] E. J. Barezi, P. Momeni, I. Wood, and P. Fung, "Modality-based factorization for multimodal fusion," in *Proc. RepL4NLP, ACL*, 2018, pp. 260–269.

[25] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 12 113–12 122.

[26] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Assoc. Comput. Linguistics*, Jul. 2019, pp. 6558–6569.

[27] Y. H. H. Tsai, P. P. Liang, A. Zadeh, L. P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Representations*, 2019.

[28] P. P. Liang, Y. C. Lim, Y. H. Tsai, R. R. Salakhutdinov, and L.-P. Morency, "Strong and simple baselines for multimodal utterance embeddings," in *North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 2599–2609.

[29] M. Chen, S. Wang, P. P. Liang, T. Baltrusǎitis, A. Zadeh, and L. P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[30] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2225–2235.

[31] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Trans. Affective Comput.*, pp. 1–1, 2020.

[32] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L. P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Representations Workshop*, 2013, pp. 1725–1732.

[34] J. Pennington, R. Socher, and C. D. Manning, "GLOVE: Global vectors for word representation," in *Proc. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North Amer. Chapter Assoc. Comput. Linguistics*, Jun. 2019, pp. 4171–4186.

[36] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 1033–1038.

[37] V. P. Rosas, R. Mihalcea, and L. P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 38–45, May/Jun. 2013.

[38] F. Eyben, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[39] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[40] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP: A collaborative voice analysis repository for speech technologies," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.

[41] S. Ji, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern, Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[42] A. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust. Speech, Signal, Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[43] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2493–2537, 2011.

[44] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2017, pp. 1243–1252.

[45] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," *arXiv:1702.01923.*

[46] S. Bai, J. Kolter, and V. Koltun, "Trellis networks for sequence modeling," in *Proc. Int. Conf. Learn. Representations*, 2019.

[47] M. Rochan, L. Ye, and Y. Wang, "Video summarization using fully convolutional sequence networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 347–363.

[48] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Pro. Int. Conf. Acoust. Speech Signal Process.*, 2019, pp. 6875–6879.

[49] Y. Liu and D. Wang, "Divide and conquer: A deep casa approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[50] R. Sen, H.-F. Yu, and I. S. Dhillon, "Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4838–4847.

[51] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[52] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Representations*, 2016.

[53] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 2011–2023, 2020.

[54] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–14.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[56] N. R. Ke *et al.*, "Sparse attentive backtracking: Temporal creditassignment through reminding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7640–7651.

[57] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[58] A. Zadeh *et al.*, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[59] S. S. Rajagopalan, L. P. Morency, T. Baltrusǎitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 338–353.

[60] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[61] D. S. Chauhan, M. S. Akhtar, A. Ekbal, and P. Bhattacharyya, "Context-aware interactive attention for multi-modal sentiment and emotion analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2019, pp. 5651–5661.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[63] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1491–1498.

[64] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Acoustical Soc. Amer. J.*, vol. 123, 2008, Art. no. 3878.

[65] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, no. 1, 2020, pp. 164–172.

[66] J. G. Zilly, R. K. Srivastava, J. Koutník, and J. Schmidhuber, "Recurrent highway networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 4189–4198.
[67] G. Kurata, B. Ramabhadran, G. Saon, and A. Sethy, "Language modeling with highway lstm," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 244–251.
[68] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, *arXiv:1505.00387*.

Sijie Mai is currently working toward the master's degree with Sun Yat-sen University, Guangzhou, China. He has authored or coauthored papers in the ACL, AAAI and IEEE TRANSACTIONS ON MULTIMEDIA. His main research interests include natural language processing and pattern recognition, including multimodal machine learning especially multimodal affective analysis, and few or zero-shot learning.

Haifeng Hu received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2004. Since July 2009, he has been a Professor with the School of Electronics and Information Engineering, Sun Yat-sen University. He has authored or coauthored about 200 papers. His research interests include computer vision, pattern recognition, image processing, and neural computation.

Songlong Xing is currently a Postgraduate with the School of Electronics and Information Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include machine learning, pattern recognition, multimodal sentiment analysis, and image captioning.