Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing

Sijie Mai
School of Electronics and Information Technology, Sun Yat-sen University
{maisj,xingslong}@mail2.sysu.edu.cn

Haifeng Hu
School of Electronics and Information Technology, Sun Yat-sen University

Songlong Xing
School of Electronics and Information Technology, Sun Yat-sen University
huhaif@mail.sysu.edu.cn

Abstract

We propose a general strategy named 'divide, conquer and combine' for multimodal fusion. Instead of directly fusing features at holistic level, we conduct fusion hierarchically so that both local and global interactions are considered for a comprehensive interpretation of multimodal embeddings. In the 'divide' and 'conquer' stages, we conduct local fusion by exploring the interaction of a portion of the aligned feature vectors across various modalities lying within a sliding window, which ensures that each part of multimodal embeddings are explored sufficiently. On its basis, global fusion is conducted in the 'combine' stage to explore the interconnection across local interactions, via an Attentive Bi-directional Skip-connected LSTM that directly connects distant local interactions and integrates two levels of attention mechanism. In this way, local interactions can exchange information sufficiently and thus obtain an overall view of multimodal information. Our method achieves state-of-the-art performance on multimodal affective computing with higher efficiency.

1 Introduction

Multimodal machine learning, as prior research shows (Baltrušaitis et al., 2019), always yields higher performance in multimodal tasks compared to the situation where only one modality is involved. In this paper, we aim at the multimodal machine learning problem, with an emphasis on multimodal affective computing where the task is to infer human's opinion from given language, visual and acoustic modalities (Poria et al., 2017a).

Finding a feasible and effective solution to learning inter-modality dynamics has been an intriguing and important problem in multimodal learning (Baltrušaitis et al., 2019), where inter-modality dynamics represent complementary information contained in more than one involved modality to be detected and analyzed for a more accurate comprehension. For this purpose, a large body of prior work mostly treats the feature vectors of the modalities as the smallest units and fuse them at holistic level (Barezi et al., 2018; Poria et al., 2016a, 2017b; Liu et al., 2018). Typically, Zadeh et al. (2017) propose a tensor-based fusion method which fuses feature vectors of three modalities using Cartesian product. Despite the effectiveness this type of methods have achieved, they give little consideration to acknowledging the variations across different portions of a feature vector which may contain disparate aspects of information and thus fail to render the fusion procedure more specialized. Additionally, they conduct fusion within one step, which can be intractable in some scenarios where the fusion method is susceptible to high computational complexity.

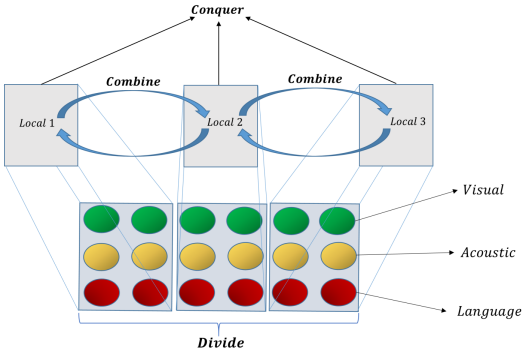

Figure 1: Schematic Diagram of our fusion strategy. Here the window size and stride are both set to 2.

Recently, Convolution Neural Networks (CNN) have achieved compelling successes in computer vision (Krizhevsky et al., 2012; Mehta et al., 2019). One of core spirits in CNN lies in the use of convolutional operation to process feature maps, which is a series of local operations with kernels sliding through the object. Inspired by it, we propose local fusion to explore local interactions in multimodal embeddings, which is in spirit similar to convolution but basically is a general strate-

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 481–492
Florence, Italy, July 28 - August 2, 2019.

gy towards multimodal fusion with multiple concrete fusion methods to choose from. Specifically, as shown in Fig. 1, we align feature vectors of three modalities to obtain multimodal embeddings and apply a sliding window to slide through them. The parallel portions of feature vectors within each window are then fused by a specific fusion method. By considering local interactions we achieve three advantages: 1) render fusion procedure more specialized since each portion of modality embeddings contains specific aspect of information intuitively; 2) assign proper weights to different portions; 3) reduce computational complexity and parameters substantially by dividing holistic fusion into multiple local ones. Many approaches can be adapted into our strategy for local fusion, and we empirically apply outer product, following (Zadeh et al., 2017). While using outer product (bilinear pooling) always brings heavy time and space complexity (Lin et al., 2015; Zadeh et al., 2017; Wu et al., 2017), we show that our method can achieve much higher efficiency.

Nonetheless, local fusion alone is not adequate for a comprehensive analysis of opinion. In fact, local interactions may contain complementary information to each other, which should be drawn upon for overall comprehension. Moreover, a small-sized sliding window may not be able to cover a complete interaction. Thus, we propose global fusion to explore interconnections of local interactions to mitigate these problems. In practice, RNN variants (Goudreau et al., 1994), especially LSTM (Hochreiter and Schmidhuber, 1997), are suitable for global fusion for their impressive power in modeling interrelations. However, in vanilla RNN architecture, only consecutive time steps are linked through hidden states, which may not be adequate for conveying information to local interactions that are far apart. Recently, some works have focused upon introducing residual learning into RNNs (Tao and Liu, 2018; Wang and Wang, 2018; Wang and Tian, 2016; He et al., 2016). Motivated by these efforts, we propose an Attentive Bi-directional Skip-connected LSTM (ABS-LSTM) that introduces bidirectional skip connection of memory cells and hidden states into LSTM, which is effective in ensuring sufficient flow of information in multi-way and handling long-term dependency problem (Bengio et al., 1994). In the transmission process of ABS-LSTM, the previous interactions are not equally correlated to the current local interaction, i.e., they vary in the amount of complementary information to be delivered. In addition, given that the local interactions, which do not contain equally valuable information, are used as input into ABS-LSTM across time steps, it is understandable that the produced states do not contribute equally to recognizing emotion. Thus, we incorporate two levels of attention mechanism into ABS-LSTM, i.e., Regional Interdependence Attention and Global Interaction Attention. The former takes effect in the process of delivering complementary information between local interactions, identifying the various correlation of previous t local interactions to the current one. The latter serves the purpose of allocating more attention to states that are more informative so as to aid a more accurate prediction.

To sum up, we propose a Hierarchical Feature Fusion Network (HFFN) for multimodal affective analysis. The main contributions are as follows:

- We propose a generic hierarchical fusion strategy, termed 'divide, conquer and combine', to explore both local and global interactions in multiple stages each focusing on different dynamics.
- Instead of conducting fusion on a holistic level, we innovate to leverage a sliding window to explore inter-modality dynamics locally. In this way, our model can take into account the variations across portions in a feature vector. Such setting also brings about an impressive bonus, i.e., significant drop in computational complexity compared to other tensor-based methods, which is proven empirically.
- We propose global fusion to obtain an overall view of multimodal embeddings via a specifically designed ABS-LSTM, in which we integrate two levels of attention mechanism: Regional Interdependence Attention and Global Interaction Attention.

2 Related Work

Previous research on affective analysis focuses on text modality(Liu and Zhang, 2012; Cambria and Hussain, 2015), which is a hot research topic in the NLP community. However, recent research suggests that information from text is not sufficient for mining opinion of humans (Poria et al., 2017a; D'Mello and Kory, 2015; Cambria, 2016), espe-

cially under the situation where sarcasm or ambiguity occurs. Nevertheless, if the accompanying information such as speaker's facial expressions and tones are presented, it would be much easier to figure out the real sentiment (Pham et al., 2019, 2018). Therefore, multimodal affective analysis has attracted increasing attention, whose major challenge is how to fuse features from various modalities. Earlier feature fusion strategies can be roughly categorized into feature-level and decision-level fusion. The former seeks to extract features of various modalities and conduct fusion at input level, by mapping them into the same embedding space simply using concatenation (Wollmer et al., 2013; Rozgic et al., 2012; Morency et al., 2011; Poria et al., 2016a, 2017b; Gu et al., 2017). The latter, by contrast, draws tentative decisions based on involved modalities separately and weighted-average the decisions, realizing cross-modal fusion (Wu and Liang, 2010; Nojavanasghari et al., 2016; Zadeh et al., 2016a; Wang et al., 2016). These two lines of work do not effectively model cross-modal or modality-specific dynamics (Zadeh et al., 2017).

Recently, word-level fusion methods have received substantial research attention and been widely acknowledged for effective exploration of time-dependent interactions (Wang et al., 2019; Zadeh et al., 2018a,b,c; Gu et al., 2018a; Rajagopalan et al., 2016). For example, Chen et al. (2017) and Gu et al. (2018b) leverage word-level alignment between modalities and explore time-restricted cross-modal dynamics. Liang et al. (2018a) propose Recurrent Multistage Fusion Network (RMFN) which decomposes multimodal fusion into three stages and uses LSTM to perform local fusion. RMFN adopts the strategy of 'divide and conquer', while our method extends it by adding 'combine' part to learn the relations between local interactions. Liang et al. (2018b) conducts emotion recognition using local-global emotion intensity rankings and Bayesian ranking algorithms. However, the 'local' and 'global' here is totally different from ours, with its 'local' referring to an utterance of a video while our 'local' represents a feature chunk of an utterance.

Tensor fusion has also become increasingly popular. Tensor Fusion Network (TFN) (Zadeh et al., 2017) adopts outer product to conduct fusion at holistic level, which is later extended by Liu et al. (2018) and Barezi et al. (2018) that try to improve efficiency and reduce redundant information by decomposing weights of high-dimensional fused tensors. HFFN mainly applies outer product as local fusion methods, and it improves efficiency by dividing modality embeddings into multiple local chunks before fusion which prevents high-dimensional fused tensor from being created. Actually, HFFN can adopt any fusion strategy in local fusion stage other than only outer product, showing high flexibility and applicability.

3 Algorithm

As shown in Fig. 2, HFFN consists of: 1) Local Fusion Module (**LFM**) for fusing features of different modalities at every local chunk; 2) Global Fusion Module (**GFM**) for exploring global inter-modality dynamics; 3) Emotion Inference Module (**EIM**) for obtaining the predicted emotion.

3.1 Divide and Conquer: Local Fusion

At the local fusion stage, we apply a sliding window that slides through the aligned feature vectors synchronously. At each step of operation, local fusion is conducted for the portions of feature vectors within the window. In this way, features across all modalities at the same window are able to fully interact with one another to obtain locally confined interactions in a more specialized way.

Assume that we have three modalities' feature vectors as input, namely language $\boldsymbol{l} \in \mathbb{R}^k$, visual $\boldsymbol{v} \in \mathbb{R}^k$ and acoustic $\boldsymbol{a} \in \mathbb{R}^k$ (we only consider the situation where all modalities share the same feature length k since they can be easily mapped into the same embedding space via some transformations). In 'divide' stage, we align these feature vectors to form the multimodal embedding $\mathbf{M} \in \mathbb{R}^{3\times k}$ and leverage a sliding window of size $3 \times d$ to explore inter-modality dynamics. Through the sliding window, each feature vector can be seen as segmented into multiple portions, each termed as a local portion. The segmentation procedure for feature vector of one modality is equivalent to:

$$\boldsymbol{m}_i=[m_{s\cdot(i-1)+1}, m_{s\cdot(i-1)+2}, ..., m_{s\cdot(i-1)+d}] \quad (1)$$

where $m \in \{l, v, a\}$ is the modality m, d is the window size, s is the stride and $\boldsymbol{m}_i$ denotes the i^{th} local portion of modality m ($i \in [1, n]$, n is the number of local portions for each modality). Obviously, for each modality, we have $n = \frac{k-d}{s} + 1$ local portions in total, provided that $k-d$ is divisible by s. Otherwise the feature vectors are padded with $0s$ to guarantee divisibility and in this case we

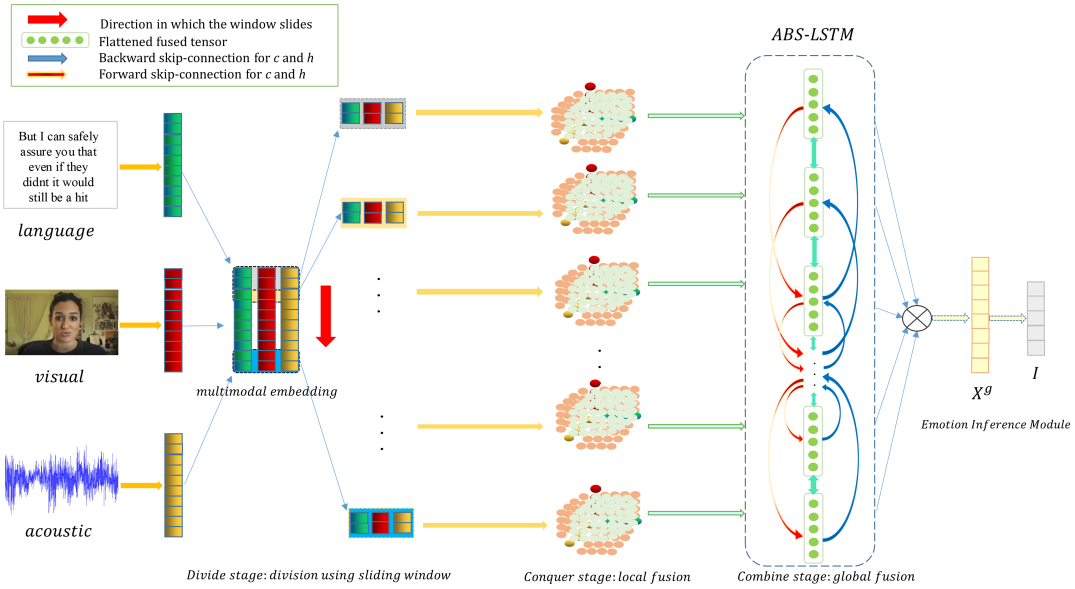

Figure 2: The Detailed Structure of HFFN.

have $n = \lfloor \frac{k-d}{s} \rfloor + 2$ local portions. In practice, both d and s can be set freely (see Section 4.3.5). For descriptive convenience we also term the parallel local portions corresponding to all modalities within the sliding window as a local chunk.

Many fusion methods can be chosen for fusing features within each local chunk to explore inter-modality dynamics in 'conquer' stage. In practice, we apply outer product for it provides the best results in our experiments. Firstly, each local portion is padded with *1s* to retain interactions of any subset of modalities as in (Zadeh et al., 2017):

$$\boldsymbol{m}_{i'} = [\boldsymbol{m}_i,\ 1],\ \ 1 \le i \le n,\ m \in \{l, v, a\} \quad (2)$$

Then we perform outer product from feature vectors padded with *1s*, defined as (Liu et al., 2018):

$$\boldsymbol{X}_i^f = \bigotimes_m \boldsymbol{m}_{i'},\ \ \boldsymbol{m}_{i'} \in \mathbb{R}^{d+1} \quad (3)$$

where $\bigotimes$ denotes tensor outer product of a set of vectors. The final local fused tensor for i^{th} local chunk is $\boldsymbol{X}_i^f \in \mathbb{R}^{(d+1)^3}$ which represents the i^{th} local interaction. We group all n local fused tensors to obtain the overall fused tensor sequence: $\boldsymbol{X}^f = [\boldsymbol{X}_1^f;\ \boldsymbol{X}_2^f;\ \ldots\ ;\ \boldsymbol{X}_n^f] \in \mathbb{R}^{n\times(d+1)^3}$. A tensor fusion diagram is shown in LFM module of Fig. 2. Compared with other models adopting outer product (Zadeh et al., 2017), our model achieves a marked improvement in efficiency by dividing holistic tensor fusion into multiple local ones, which is shown in Section 4.3.3. Actually, we can apply other fusion methods that are suitable for local information extraction, which demonstrates the broad applicability of our strategy and is left for future work.

3.2 Combine: Global Fusion

In the 'combine' stage, we model global interactions by exploring interconnections (complementary information) and context-dependency across local fused tensors to obtain an overall interpretation of interactions comprehensively. In addition, the limited and fixed size of sliding window may lead to division of the complete process of expressing emotion into different local portions, in which case sufficient flow of information between local chunks is warranted to compensate for this problem. Therefore, we design ABS-LSTM, an RNN variant, to make sense of the cross-modality dynamics from an integral perspective. In ABS-LSTM, we introduce bidirectional residual connection of memory cells and hidden states as well as integrate attention mechanisms to transmit information and learn overall representations more effectively, as shown in Fig. 2. Now that we obtain the local fused tensor sequence $\boldsymbol{X}^f$ in LFM, global interaction learning can be expressed as:

$$\boldsymbol{X}^g = \textit{ABS-LSTM}(\boldsymbol{X}^f) \quad (4)$$

where *ABS-LSTM* is activated by $tanh$ nonlinear function, $\boldsymbol{X}^g = [\boldsymbol{X}_1^g;\ \boldsymbol{X}_2^g; \ldots; \boldsymbol{X}_n^g] \in \mathbb{R}^{n\times 2o}$ is the global fused tensor sequence, and $2o$ is the dimensionality of ABS-LSTM's output. A detailed illustration of ABS-LSTM is shown below.

3.2.1 ABS-LSTM

ABS-LSTM is specifically designed for modeling the interconnections of local fused tensors to distill complementary information. Since local in-

teractions within a certain distance range are mutually correlated, it is necessary for ABS-LSTM to operate in a bidirectional way. As opposed to conventional bidirectional RNNs, ABS-LSTM has a set of identical parameters for both forward/backward passes which ensures a smaller number of parameters. Further, ABS-LSTM directly connects the current interaction with its several neighbors so that information can be sufficiently exchanged. Given its ability to bidirectionally transmit information in multiple connections, it is powerful in modeling long-term dependency, which is crucial for long sequences.

Firstly we illustrate the pipeline of ABS-LSTM in forward pass stage. Assume that t previous local interactions are directly connected to the current one (t is set to 3 in our experiment), it is beneficial to identify the various correlation between previous t interactions and the current interaction. To this end, we integrate Regional Interdependence Attention (RIA) into ABS-LSTM, so that previous local interactions containing more complementary information to the current one are given more importance in the information transmission process. The equations for previous information fusion of cells and states for l^{th} interaction in forward pass are as follows:

$$\begin{aligned} \boldsymbol{s}_{c_{l-i}} &= tanh(\boldsymbol{W}_c(\overrightarrow{\boldsymbol{c}}_{l-i} \oplus \boldsymbol{X}_l^f)), 1 \le i \le t \\ \boldsymbol{s}_{h_{l-i}} &= tanh(\boldsymbol{W}_h(\overrightarrow{\boldsymbol{h}}_{l-i} \oplus \boldsymbol{X}_l^f)), 1 \le i \le t \end{aligned} \tag{5}$$

$$\begin{aligned} \boldsymbol{s}_c &= [\|\boldsymbol{s}_{c_{l-t}}\|_2, \|\boldsymbol{s}_{c_{l-t+1}}\|_2, ..., \|\boldsymbol{s}_{c_{l-1}}\|_2], \\ \boldsymbol{s}_h &= [\|\boldsymbol{s}_{h_{l-t}}\|_2, \|\boldsymbol{s}_{h_{l-t+1}}\|_2, ..., \|\boldsymbol{s}_{h_{l-1}}\|_2] \end{aligned} \tag{6}$$

$$\boldsymbol{\gamma}_c = softmax(\boldsymbol{s}_c), \boldsymbol{\gamma}_h = softmax(\boldsymbol{s}_h) \tag{7}$$

$$\tilde{\boldsymbol{c}}_l = \boldsymbol{a}(\sum_{i=1}^{t} \gamma_{c_{l-i}} \overrightarrow{\boldsymbol{c}}_{l-i}), \tilde{\boldsymbol{h}}_l = \boldsymbol{a}(\sum_{i=1}^{t} \gamma_{h_{l-i}} \overrightarrow{\boldsymbol{h}}_{l-i}) \tag{8}$$

where $\bigoplus$ denotes vector concatenation and $\boldsymbol{W}_h, \boldsymbol{W}_c \in \mathbb{R}^{o \times (o+(d+1)^3)}$ are parameter matrices that determine the importance of previous cells $\overrightarrow{\boldsymbol{c}}_{l-i}$ and states $\overrightarrow{\boldsymbol{c}}_{l-i}$, respectively. Eq. 5 maps the cell and state at the $(l-i)^{th}$ time step into two o-dimensional vectors respectively. Instead of merely using $\overrightarrow{\boldsymbol{c}}_{l-i}$ or $\overrightarrow{\boldsymbol{h}}_{l-i}$ to obtain their importance towards local interaction at current time step, we also utilize current time step's input $\boldsymbol{X}_l^f$ to reflect the correlation between the cell and states of $(l-i)^{th}$ interaction and current l^{th} time step's input, which provides a better measurement of attention score by learning inter-dependency correlation between interactions. We take the 2-norm of each vector in Eq. 6 as the importance score of each previous cell and state and then form a t-dimensional importance score vector for all states and cells, respectively. In Eq. 7 we use $softmax$ layer to normalize both vectors and obtain the final attention scores, which, according to Eq. 8, are used as weights for the combination of previous t local interactions. The function $\boldsymbol{a}$ in Eq. 8 is a nonlinear activation function that helps to improve expressive power of ABS-LSTM, which we empirically choose $ReLU$. Overall, Eq. 5 to Eq. 8 realize transmission of information from previous multiple local interactions to the current one, using the first level of attention mechanism, i.e., RIA, which is able to properly distribute attention across the previous t local interactions to focus on the ones that contain information most relevant to the current local interaction.

After the combination of previous information, we further define:

$$\boldsymbol{f}_l = \sigma(\boldsymbol{W}_{f_1} \boldsymbol{X}_l^f + \boldsymbol{W}_{f_2} \tilde{\boldsymbol{h}}_l) \tag{9}$$

$$\boldsymbol{i}_l = \sigma(\boldsymbol{W}_{i_1} \boldsymbol{X}_l^f + \boldsymbol{W}_{i_2} \tilde{\boldsymbol{h}}_l) \tag{10}$$

$$\overrightarrow{\boldsymbol{c}}_l = \boldsymbol{f}_l \odot \tilde{\boldsymbol{c}}_l + \boldsymbol{i}_l \odot tanh(\boldsymbol{W}_{m_1} \boldsymbol{X}_l^f + \boldsymbol{W}_{m_2} \tilde{\boldsymbol{h}}_l) \tag{11}$$

$$\overrightarrow{\boldsymbol{h}}_l = \sigma(\boldsymbol{W}_{o_1} \boldsymbol{X}_l^f + \boldsymbol{W}_{o_2} \tilde{\boldsymbol{h}}_l) \odot tanh(\overrightarrow{\boldsymbol{c}}_l) \tag{12}$$

where σ denotes $sigmoid$ function. Eq. 9 - 12 denote the routine procedure of LSTM except that $\overrightarrow{\boldsymbol{h}}_{l-1}$ and $\overrightarrow{\boldsymbol{c}}_{l-1}$ are replaced with $\tilde{\boldsymbol{h}}_l$ and $\tilde{\boldsymbol{c}}_l$, respectively. The output of l^{th} time step in forward pass stage is $\overrightarrow{\boldsymbol{h}}_l$ $(1 \le l \le n)$. To make ABS-LSTM bidirectional, in backward pass stage, we reverse input $\boldsymbol{X}^f$ so that the last interaction arrives in first place and again feed it into Eq. 5 - 12, whose output becomes $\overleftarrow{\boldsymbol{h}}_l$. The output of ABS-LSTM at l^{th} time step is: $\boldsymbol{h}_l = \overleftarrow{\boldsymbol{h}}_l \bigoplus \overrightarrow{\boldsymbol{h}}_l \in \mathbb{R}^{2o}$

Global Interaction Attention (GIA): Inherently, LSTM has the capability to 'memorize', and uses the memory to sequentially model long-term dependency. Thus, the hidden states output by ABS-LSTM synthesize the information from current time step's input interaction and that from previous input, respectively. In this sense, at each time step new information is processed and previous information still exists but is 'diluted' in the hidden state (due to the forget gate). Therefore, as some local interactions that are more informative, e.g. revealing a sharp tone or sheer alteration of facial expressions, are input to ABS-LSTM, the produced states should be given more importance

over others since they have just synthesized an informative interaction and not yet been 'diluted'. Hence, it is justifiable to employ a specifically designed attention mechanism, termed Global Interaction Attention (GIA), to properly assign importance across states. GIA is formulated as follows:

$$\boldsymbol{\omega}_h = ReLU(\boldsymbol{W}_{h'}\boldsymbol{h}_l + \mathbf{b}_{h'}) \quad (13)$$

$$\boldsymbol{\omega}_x = ReLU(\boldsymbol{W}_x\boldsymbol{X}_l^f + \mathbf{b}_x) \quad (14)$$

$$\boldsymbol{h}_l^a = tanh((\boldsymbol{W}_{h_2}\boldsymbol{\omega}_h)\cdot\boldsymbol{h}_l + \boldsymbol{W}_{x_2}\boldsymbol{\omega}_x) \quad (15)$$

where $\boldsymbol{W}_{h'} \in \mathbb{R}^{o\times 2o}$ and $\boldsymbol{W}_x \in \mathbb{R}^{o\times(d+1)^3}$ are two parameter matrices and $\mathbf{b}_{h'}, \mathbf{b}_x \in \mathbb{R}^o$ are two bias vectors to be learned. $\boldsymbol{W}_{h_2}, \boldsymbol{W}_{x_2} \in \mathbb{R}^{1\times o}$ are two parameter matrices that determine final importance scores. Through affine transforms and nonlinearities in Eq. 13 and Eq. 14, the l^{th} state $\boldsymbol{h}_l$ and the corresponding input $\boldsymbol{X}_l^f$ are embedded into two o-dimensional vectors $\boldsymbol{\omega}_h$ and $\boldsymbol{\omega}_x$ that contain information regarding importance of l^{th} state and local interaction, respectively. In Eq. 15, $\boldsymbol{W}_{h_2}$ and $\boldsymbol{\omega}_h$ first form a scalar via matrix multiplication, which reflects the importance of the l^{th} hidden state to be used as its weight. Meanwhile, we pre-multiply $\boldsymbol{\omega}_x$ by $\boldsymbol{W}_{x_2}$ and obtain a scalar to be added to each entry of weighted state, which functions as a bias containing input information. By this means, the attended state at current time step is able to focus more on the information from current interaction instead of the previous ones. Considering that $\boldsymbol{X}_l^f$ and $\boldsymbol{h}_l$ are two intrinsically disparate sources of information, we only formulate the impact of $\boldsymbol{X}_l^f$ as a scalar that biases the state, rather than as a vector which has much more complex influence to the state and empirically degrades performance. In this way, if $\boldsymbol{X}_l^f$ is more important, the l^{th} attended state $\boldsymbol{h}_l^a$ will receive a more significant shift towards a higher position with respect to all high-dimensional coordinates, and thus $\boldsymbol{h}_l^a$ is more attended. In a sense, every element of the original state undergoes a transformation, with a specifically determined weight and a fixed bias across all entries. GIA enables ABS-LSTM to enhance the states of greater importance, aiding a more accurate classification. The final output of ABS-LSTM is the concatenation of attended states: $\boldsymbol{X}^g = \bigoplus_{l=1}^n \boldsymbol{h}_l^a \in \mathbb{R}^{n\times 2o}$.

3.3 Emotion Inference Module

After obtaining the global interactions, the final emotion is obtained by:

$$\mathbf{E} = \boldsymbol{f}(\mathbf{W}_{e1}\boldsymbol{X}^g + \mathbf{b}_{e1}) \quad (16)$$

$$\boldsymbol{I} = softmax(\mathbf{W}_{e2}\mathbf{E}) \quad (17)$$

where $\boldsymbol{f}$ contains a $tanh$ activation function and a dropout layer of dropout rate 0.5, $\mathbf{W}_{e1} \in \mathbb{R}^{50\times n\cdot 2o}$, $\mathbf{b}_{e1} \in \mathbb{R}^{50}$ and $\mathbf{W}_{e2} \in \mathbb{R}^{N\times 50}$ are the learnable parameters, and $\boldsymbol{I} \in \mathbb{R}^N$ is the final emotion inference (N is the number of categories).

4 Experiments

4.1 datasets

CMU-MOSI (Zadeh et al., 2016b) includes 93 videos with each video padded to 62 utterances. We consider positive and negative sentiments in our paper. We use 49 videos for training, 13 for validation and 31 for testing. **CMU-MOSEI** (Zadeh et al., 2018c) has 2928 videos, and each video is padded to 98 utterances. Each utterance has been scored on two perspectives: sentiment intensity (ranges between [-3, 3]) and emotion (six classes). We consider positive, negative and neutral sentiments in the paper. We utilize 1800, 450 and 678 videos respectively for training, validation and testing. **IEMOCAP** (Busso et al., 2008) contains 151 videos and each video has at most 110 utterances. IEMOCAP contains following labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise and other. We take the first four emotions so as to compare with previous models. The training, validation and testing sets contain 96, 24 and 31 videos respectively.

4.2 Experimental details

HFFN is implemented using the framework of Keras, with tensorflow as backend. The input dimensionality k for CMU-MOSI and CMU-MOSEI datasets is 50, while for IEMOCAP, k is set to 100. We use RMSprop for optimizing the network, with cosine proximity as objective function. The output dimension $2o$ of ABS-LSTM is set to 6 for CMU-MOSI and CMU-MOSEI but 2 for IEMOCAP. Note that ABS-LSTM is activated by $tanh$ and followed by a dropout layer.

For feature pre-extraction, our setting on CMU-MOSI and IEMOCAP datasets are identical to that in (Poria et al., 2017b)[1]. The features are extracted from each utterances separately. For **language feature**, a text-CNN is applied. Each word is first embedded into a vector using word2vec tool (Mikolov et al., 2013). Then

[1]https://github.com/soujanyaporia/multimodal-sentiment-analysis

the vectorized representations for all words in an utterance are concatenated, which afterwards is processed by CNNs (Karpathy et al., 2014). For **acoustic feature**, an open-source tool openSMILE (Eyben, 2010) is utilized to generate high dimensional vectors comprised of low-level descriptors (LLD). 3D-CNN (Ji et al., 2013) is applied for **visual feature pre-extraction**. It learns relevant features from each frame and the alterations across consecutive frames. By contrast, on CMU-MOSEI dataset we follow the setting as in (Zadeh et al., 2017; Liu et al., 2018)[2]. GloVe (Pennington et al., 2014), Facet (iMotions, 2017) and COVAREP (Degottex et al., 2014) are applied for extracting language, visual and acoustic features respectively. Word-level alignment is performed using P2FA (Yuan and Liberman, 2008) across modalities. Eventually the unimodal features are generated as the average of their feature values over word time interval (Chen et al., 2017).

Subsequent to pre-extraction, similar to BC-LSTM (Poria et al., 2017b), we devise a Unimodal Feature Extraction Network (UFEN): $\mathbb{R}^{u\times d_j} \rightarrow \mathbb{R}^{u\times k}$, which consists of a bidirectional LSTM layer followed by a fully connected (FC) layer, for each separate modality. Here, u denotes the number of utterances that constitute a video and d_j is the dimensionality of raw feature vector for j^{th} modality. Through UFEN, feature vectors of all modalities are mapped into the same embedding space (have the same dimensionality k). UFEN for each modality, is individually trained followed by a FC layer: $\mathbb{R}^k \rightarrow \mathbb{R}^N$ using Adadelta (Zeiler, 2012) as optimizer and with categorical cross-entropy as loss function. The precessed feature vectors of each utterance will be sent into HFFN.

4.3 Results and Discussions

4.3.1 Comparison with Baselines

We compare HFFN with following multimodal algorithms: RMFN (Liang et al., 2018a), MFN (Zadeh et al., 2018a), MCTN (Pham et al., 2019), BC-LSTM (Poria et al., 2017b), TFN (Zadeh et al., 2017), MARN (Zadeh et al., 2018b), LMF (Liu et al., 2018), MFM (Tsai et al., 2019), MRRF (Barezi et al., 2018), FAF (Gu et al., 2018b), RAVEN (Wang et al., 2019), GMFN (Zadeh et al., 2018c), Memn2n (Sukhbaatar et al., 2015), MMB2 (Liang et al., 2019), CHFusion (Majumder et al., 2018), SVM Trees (Rozgic et al., 2012), CMN (Hazarika et al., 2018), C-MKL (Poria et al., 2016b) and CAT-LSTM (Poria et al., 2017c).

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

Methods	Acc	F1 Score
BC-LSTM	77.9	78.1
CAT-LSTM	76.6	76.2
MFN	77.4	77.3
FAF	76.5	76.8
RAVEN	78.0	-
CHFusion	80.0	-
MMB2	75.2	75.1
MFM	77.4	77.3
RMFN	78.4	78.0
MCTN	79.3	79.1
GMFN	77.7	77.7
TFN	74.6	74.5
LMF	76.4	75.7
MRRF	73.0	73.1
HFFN($d, s = 2, 2$)	**80.19**	**80.34**

Table 1: Performance on CMU-MOSI dataset.

Models	Acc	F1 score
TFN	59.40	57.33
LMF	60.27	53.87
CHFusion	58.45	56.90
BC-LSTM	**60.77**	59.04
CAT-LSTM	60.72	58.83
HFFN($d, s = 2, 2$)	60.37	**59.07**

Table 3: Performance on CMU-MOSEI dataset.

As presented in Table 1, HFFN shows improvement over typical approaches, setting new state-of-the-art record. Compared with the tensor fusion approaches TFN (Zadeh et al., 2017), MRRF (Barezi et al., 2018) and LMF (Liu et al., 2018), HFFN achieves improvement by about 4%, which demonstrates its superiority. It is reasonable because these methods conduct tensor fusion at holistic level and ignore modeling local interactions, while ours has a well-designed LFM module. Compared to the word-level fusion approaches RAVEN (Wang et al., 2019), RMFN (Liang et al., 2018a) and FAF (Gu et al., 2018b), etc., HFFN achieves improvement by about 2%. We argue that it is because they ignore explicitly connecting locally-constrained interactions to obtain a general view of multimodal signals, while we explore global interactions by applying ABS-LSTM.

The results on IEMOCAP and CMU-MOSEI datasets are shown in Table 2 and Table 3, respectively. We can conclude from Table 2 that HFFN achieves consistent improvements on accuracy and F1 score in IEMOCAP 4-way and individual emotion recognition tasks compared with other methods. Specifically, HFFN outperforms other methods by a significant margin on the recognition of Angry and Neutral emotions. For CMU-MOSEI dataset, as shown in Table 3, the accuracy of HFFN is lower than that of BC-LSTM and CAT-LSTM, but it achieves the highest F1 s-

	IEMOCAP (Individual)				IEMOCAP (4-way)		
Models	F1-Angry	F1-Happy	F1-Sad	F1-Neutral	Models	Acc	F1
MFM	86.7	85.8	86.1	68.1	C-MKL	74.1	-
MARN	84.2	83.6	81.2	66.7	CHFusion	76.5	76.8
MFN	83.7	84.0	82.1	69.2	SVM Trees	67.4	-
RMFN	84.6	85.8	82.9	69.1	BC-LSTM	77.57	77.80
TFN	84.2	83.6	82.8	65.4	CAT-LSTM	80.47	80.27
MRRF	86.0	85.6	85.8	67.9	Memn2n	75.08	-
GMFN	85.5	84.2	83.0	68.9	CMN	77.62	-
RAVEN	86.7	85.8	83.1	69.3	TFN	75.83	75.99
LMF	89.0	85.8	85.9	71.7	LMF	76.32	76.49
HFFN$(d, s = 2, 2)$	**94.31**	**88.65**	**86.24**	**76.24**	HFFN$(d, s = 2, 2)$	**82.37**	**82.42**

Table 2: Performance of HFFN on IEMOCAP dataset. Here F1- means F1 score.

	CMU-MOSI		IEMOCAP	
Methods	Acc	F1	Acc	F1
L	78.59	78.52	81.46	81.54
A	48.14	48.30	38.08	38.17
V	56.97	57.48	34.52	29.15
L+A	78.06	78.29	80.38	80.60
L+V	79.39	79.38	80.05	80.26
A+V	55.17	55.76	55.17	55.79
L+A+V	**80.19**	**80.34**	**82.37**	**82.42**

Table 4: Unimodal, Bimodal and Trimodal Results of HFFN. Here, L, A and V denotes language, acoustic and visual modalities, respectively.

core with slight margin. HFFN still achieves state-of-the-art performance on these two datasets.

4.3.2 Discussion on Modality Importance

To explore the underlying information of each modality, we carry out an experiment to compare the performance among unimodal, bimodal and trimodal models. For unimodal models, we can infer from Table 4 that language modality is the most predictive for emotion prediction, outperforming acoustic and visual modalities with significant margin. When coupled with acoustic and visual modalities, the trimodal HFFN performs best, whose result is $1\% \sim 2\%$ better than the language-HFFN, indicting that acoustic and visual modalities actually play auxiliary roles while language is dominant. However, in our model, when conducting outer product, all three modalities are treated equally, which is probably not the optimal choice. In the future, we aim to develop a fusion technique paying more attention to the language modality, while the other two modalities only serve as accessory sources of information.

Interestingly, the bimodal HFFNs do not necessarily outperform the language-HFFN. Contrarily, sometimes it even lowers the performance when language is combining with acoustic or visual modality. Nevertheless, when three modalities are available, the performance is undoubtedly the best. It indicates that a great deal of information hidden in a single modality can be interpreted only by combining all the three modalities.

Methods	FLOPs	Number of Parameters
BC-LSTM	1,322,024	1,383,902
TFN	8,491,845	4,245,986
HFFN	**16,665**	**8,301**

Table 5: Comparison of Efficiency.

4.3.3 Comparative Analysis on Efficiency

Contrast experiments are conducted to analyze the efficiency of TFN (Zadeh et al., 2017), BC-LSTM (Poria et al., 2017b) [3] and HFFN. We compare the number of parameters and FLOPs after fusion (the FLOPs index is used to measure time complexity), and the inputs for all methods are the same to make a fair comparison. The trainable layers in TFN include two FC layers of 32 $ReLU$ activation units and a decision layer: $\mathbb{R}^{32} \rightarrow \mathbb{R}^2$. We adopt this setting to match the code released by the authors [4]. BC-LSTM's trainable layers contain a bidirectional LSTM with input and output dimension being $3 \cdot 50$ and 600 respectively, and two FC layers of 500 and 2 units respectively.

Table 5 shows that in terms of the number of parameters, TFN is around 511 times larger than our HFFN, even under the situation where we adopt a more complex module after tensor fusion, demonstrating the high efficiency of HFFN. Note that if TFN adopts the original setting as stated in (Zadeh et al., 2017) where the FC layers have 128 units, it would even have more parameters than our version of TFN. Compared to BC-LSTM, HFFN has about 166 times fewer parameters and the FLOPs of HFFN is over 79 times fewer than that of BC-LSTM. Moreover, BC-LSTM is over 6 times faster than TFN in time complexity measured by FLOPs and the number of parameters is over 3 times smaller. These results demonstrate that outer product applied in TFN results in heavy computational complexity and a substantial number of param-

[3] https://github.com/soujanyaporia/multimodal-sentiment-analysis

[4] https://github.com/Justin1904/TensorFusionNetworks

	Acc	F1 score
Bidirectional LSTM	79.65	79.77
LSTM	78.86	78.97
ABS-LSTM(no attention)	79.12	79.22
ABS-LSTM(RIA)	79.39	79.54
ABS-LSTM(GIA)	79.39	79.47
ABS-LSTM(RIA+GIA)	**80.19**	**80.34**

Table 6: Discussion on LSTM Variants.

eters compared with other methods such as BC-LSTM, while HFFN can avoid these two problems and is even more efficient than other approaches adopting low-complexity fusion methods.

4.3.4 Discussion on Global Fusion

To demonstrate the superiority of ABS-LSTM on learning global interactions and the impact of the proposed attention mechanism, we conduct an experiment to compare the performance of model under different settings of global fusion. We can infer from Table 6 that ABS-LSTM reaches best results among all tested LSTM variants. Besides, vanilla LSTM achieves lowest performance, showing the necessity of delivering information bidirectionally. Bidirectional LSTM slightly outperforms no-attention variant of ABS-LSTM, possibly due to the use of two sets of independent learnable parameters for forward and backward passes, respectively, which allows more flexibility. However, as ABS-LSTM with attention outperforms bidirectional LSTM, it demonstrates the efficacy of ABS-LSTM.

In terms of the effectiveness of attention mechanisms, interestingly, both RIA and GIA, when used alone, only bring about slight improvement (0.2%∼0.3%) compared to the no-attention version of ABS-LSTM. However, it further boosts the performance when RIA and GIA are concurrently used, achieving more improvement than that caused by RIA and GIA alone added together. This shows some potential positive link between the two levels of attention mechanism. Specifically, RIA can provide more refined information during transmission between local interactions, so that the output states to be processed by GIA are more focused on useful information and freer of noise, maximizing the effect of GIA.

4.3.5 Discussion on Sliding Window

To investigate the influence of the size d and the stride s of sliding window on learning local interactions, we conduct experiments on IEMOCAP where s changes incrementally from 1 to 10 and d takes on four values, namely 1, 2, 5 and 10. The

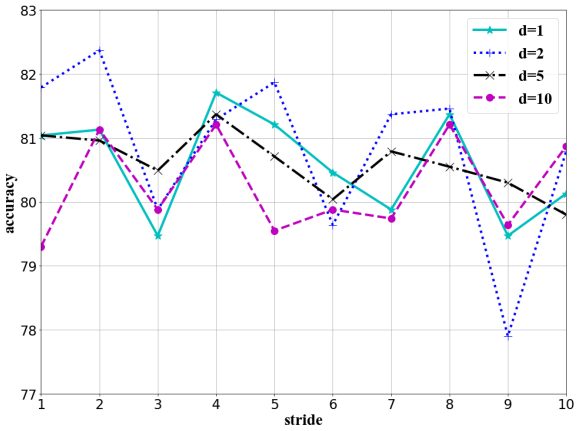

Figure 3: Influence of window size d and stride s.

results are shown in Fig. 3. It can be observed that for all values of d, the accuracy fluctuates within a limited range as the stride s changes incrementally, showing robustness with respect to the stride. Overall, the model fares best when d is set to 2, demonstrating that a moderate size of sliding window is important for ensuring high performance. We conjecture that the reason behind the decline in performance when d is assigned an overly large value (greater than 2), is that the effect of local fusion is lessened, leading to less specialized exploration of feature portions. This in turn verifies the central importance of local fusion in our strategy. In addition, an unreasonably small d may lead to disintegration of the feature correlation that could be capitalized on and scatter complete information, thus hurting overall performance. Furthermore, it is surprising that when the stride s is greater than d (some dimensions of feature vectors are left out in local fusion), the accuracy does not significantly suffer. This shows that there may be a deal of redundant information in the feature vectors, implying that more advanced extraction techniques are needed for more refined representations, which we will explore as part of future work.

5 Conclusion

We propose an efficient and effective framework HFFN that adopts a novel fusion strategy called 'divide, conquer and combine'. HFFN learns local interactions at each local chunk and explores global interactions by conveying information across local interactions using ABS-LSTM that integrates two levels of attention mechanism. Our fusion strategy is generic for other concrete fusion methods. In future work, we intend to explore multiple local fusion methods within our framework.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Elham J. Barezi, Peyman Momeni, Ian Wood, and Pascale Fung. 2018. Modality-based factorization for multimodal fusion. *CoRR*, abs/1811.12624.

Y Bengio, P Simard, and P Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

E. Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(02):102–107.

Erik Cambria and Amir Hussain. 2015. Senticnet. *Sentic Computing*, pages 23–71.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrusăitis, Amir Zadeh, and Louis Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. *19th ACM International Conference on Multimodal Interaction (ICMI'17)*.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep: A collaborative voice analysis repository for speech technologies. In *ICASSP*.

Sidney K. D'Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *Acm Computing Surveys*, 47(3):1–36.

Florian Eyben. 2010. Opensmile:the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pages 1459–1462.

M W Goudreau, C L Giles, S T Chakradhar, and . Chen, D. 1994. First-order versus second-order single-layer recurrent neural networks. *IEEE Transactions on Neural Networks*, 5(3):511–513.

Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech intention classification with multimodal deep learning. *Proceedings of Canadian Conference on Artificial Intelligence*.

Yue Gu, Xinyu Li, Kaixiang Huang, Shiyu Fu, Kangning Yang, Shuhong Chen, Moliang Zhou, and Ivan Marsic. 2018a. Human conversation analysis using attentive multimodal networks with hierarchical encoder-decoder. In *ACM Multimedia*.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018b. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. *ACL*.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. pages 2122–2132.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

iMotions. 2017. Facial expression analysis.

S. Ji, M. Yang, and K. Yu. 2013. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei Fei Li. 2014. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition*, pages 1725–1732.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*.

Paul Pu Liang, Yao Chong Lim, Y. H. Tsai, Ruslan R. Salakhutdinov, and Louis-Philippe Morency. 2019. Strong and simple baselines for multimodal utterance embeddings. In *NAACL*.

Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis Philippe Morency. 2018a. Multimodal language analysis with recurrent multistage fusion. *EMNLP*.

Paul Pu Liang, Amir Zadeh, and Louis Philippe Morency. 2018b. Multimodal local-global ranking fusion for emotion recognition. *2018 International Conference on Multimodal Interaction (ICMI'18)*, pages 472–476.

Tsung Yu Lin, Aruni Roychowdhury, and Subhransu Maji. 2015. Bilinear cnn models for fine-grained visual recognition. *ICCV*, pages 1449–1457.

Bing Liu and Lei Zhang. 2012. *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA.

Zhun Liu, Ying Shen, Paul Pu Liang, Amir Zadeh, and Louis Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *ACL*, pages 2247–2256.

N Majumder, D Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*.

Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. 2019. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Louis Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *International Conference on Multimodal Interfaces*, pages 169–176.

Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, and Louis Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of ACM International Conference on Multimodal Interaction*, pages 284–288.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis Philippe Morency, and Poczos Barnabas. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. *AAAI*.

Hai Pham, Manzini Thomos, Liang Paul Pu, and Poczos Barnabas. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *ACL 2018 Grand Challenge and Workshop on Human Multimodal Language*.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017a. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *ACL*, pages 873–883.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis Philippe Morency. 2017c. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016a. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *Proceedings of IEEE International Conference on Data Mining (ICDM)*, pages 439–448.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016b. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *IEEE International Conference on Data Mining*, pages 439–448.

Shyam Sundar Rajagopalan, Louis Philippe Morency, Tadas BaltrusÁitis, and Roland Goecke. 2016. Extending long short-term memory for multi-view structured learning. *ECCV*, pages 338–353.

V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Signal and Information Processing Association Summit and Conference*, pages 1–4.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Fei Tao and Gang Liu. 2018. Advanced lstm: A study about better time dependency modeling in emotion recognition. *Proceedings of Acoustics, Speech and Signal Processing (ICASSP)*.

Yao Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning factorized multimodal representations. *ICLR*.

Cheng Wang and Cheng Wang. 2018. Rra: Recurrent residual attention for sequence learning. *AAAI*.

Haohan Wang, Aaksha Meghawat, Louis Philippe Morency, and Eric P Xing. 2016. Select-additive learning: Improving generalization in multimodal sentiment analysis. pages 949–954.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, and Amir Zadeh. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. *AAAI*.

Yiren Wang and Fei Tian. 2016. Recurrent residual learning for sequence classification. In *EMNLP*, pages 938–943.

Martin Wollmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, and Louis Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53.

Chung Hsien Wu and Wei Bin Liang. 2010. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21.

Lin Wu, Yang Wang, Lin Wu, Yang Wang, Lin Wu, and Yang Wang. 2017. Where to focus: Deep attention-based spatially recurrent bilinear networks for fine-grained visual recognition. *arXiv Preprint: 1709.05769*.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Acoustical Society of America Journal*, 123:3878.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *EMNLP*, pages 1114–1125.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. *AAAI*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. *AAAI*.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis Philippe Morency. 2018c. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. *ACL*, pages 2236–2246.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis Philippe Morency. 2016a. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intelligent Systems*, 31(6):82–88.

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

Matthew D. Zeiler. 2012. Adadelta: An adaptive learning rate method. *preprint, arXiv:1212.5701v1*.