Dynamically Adjust Word Representations Using Unaligned Multimodal Information

Jiwei Guo
Hangzhou Dianzi University
Hangzhou, China
guojiwei@hdu.edu.cn

Jiajia Tang
Hangzhou Dianzi University
Hangzhou, China
hdutangjiajia@163.com

Weichen Dai
Hangzhou Dianzi University
Hangzhou, China
weichendai@hdu.edu.cn

Yu Ding
Netease Fuxi AI Lab
Hangzhou, China
dingyu01@corp.netease.com

Wanzeng Kong*
Hangzhou Dianzi University
Hangzhou, China
kongwanzeng@hdu.edu.cn

ABSTRACT

Multimodal Sentiment Analysis is a promising research area for modeling multiple heterogeneous modalities. Two major challenges that exist in this area are a) multimodal data is unaligned in nature due to the different sampling rates of each modality, and b) long-range dependencies between elements across modalities. These challenges increase the difficulty of conducting efficient multimodal fusion. In this work, we propose a novel end-to-end network named Cross Hyper-modality Fusion Network (CHFN). The CHFN is an interpretable Transformer-based neural model that provides an efficient framework for fusing unaligned multimodal sequences. The heart of our model is to dynamically adjust word representations in different non-verbal contexts using unaligned multimodal sequences. It is concerned with the influence of non-verbal behavioral information at the scale of the entire utterances and then integrates this influence into verbal expression. We conducted experiments on both publicly available multimodal sentiment analysis datasets CMU-MOSI and CMU-MOSEI. The experiment results demonstrate that our model surpasses state-of-the-art models. In addition, we visualize the learned interactions between language modality and non-verbal behavior information and explore the underlying dynamics of multimodal language data.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → Multimedia information systems; **Sentiment analysis**.

KEYWORDS

multimodal sentiment analysis, multimodal fusion, multimodal representations

*Wanzeng Kong is the corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548137

ACM Reference Format:
Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong*. 2022. Dynamically Adjust Word Representations Using Unaligned Multimodal Information. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548137

1 INTRODUCTION

With the rise of social networks, video sites, and online movie reviews, understanding and mining the emotional elements from these multimodal sequences [11], namely Multimodal Sentiment Analysis (MSA), is a promising research area. People convey their intentions and emotions via multimodal data in daily social activities. Such multimodal data mainly consists of three channels: language (spoken words), visual (gesture), and acoustic (voice). Different modalities in the same data segment are often complementary and actively interact, providing more robust information than just a single modality [18, 20].

For example, a person might say, "The movie is sick." This utterance is ambiguous (either positive or negative) by itself. The speaker may convey a negative emotion if he frowns while speaking. On the other hand, the same statement accompanied by a smile would be perceived positively, and a loud voice increases the sentiment to strong positive [36]. The meaning of words and sentences uttered by the speaker often varies dynamically in different non-verbal contexts [32]. The interactions between language and non-verbal behavior information (including visual and acoustic) change the perception of the language expressed sentiment. The key point to addressing existing challenges in the MSA area is how to effectively fuse meaningful information [28] from different modalities for sentiment analysis.

The raw multimodal data sources are unaligned because of diverse sampling rates for each modality [30]. In addition, the inherent asynchrony between different modalities may result in sound or subtitles that are not fully synchronized with the video. Therefore, a majority of existing works [2, 19, 25, 26, 32] require the manual alignment process, which strictly aligns visual and acoustic sequences to the resolution of textual words. Then, multimodal fusion is performed on the aligned multimodal data to complete sentiment analysis. However, the manual alignment process usually requires a huge amount of time and labor effort [14, 16], although methods have achieved excellent results. Furthermore, since a frowning face

may relate to a word pessimistically spoken in the past, the inferring long-range dependencies between elements across modalities are required in the MSA [30].

To address the above issues, some models [14, 16, 30] can perform multimodal fusion on unaligned multimodal sequences. However, the multimodal interactions in most models are limited by the number of modalities. Their interactions are a bi-modal operation that only accounts for two modalities' input at a time. When the number of modalities exceeds two, multiple operations are required to obtain all modal interactions and require a large number of parameters to preserve the original modality information. Furthermore, when they learn multimodal representations, the contribution of each modality is equal for the sentiment analysis task. But previous works [17, 19, 26, 32] indicate that language modality should play a dominant role when learning joint representations from multiple modal data sources since the language modality contains more practical information than no-verbal modalities[23], and algorithms for text analysis are well studied. If the three modalities are treated equally, the performance of the multimodal system may be weakened [17, 19].

Motivated by the above observation, we propose a Cross Hyper-modality Fusion Network (CHFN) towards multimodal fusion from unaligned modal sequences. In CHFN, the video and audio modalities are first organized into hyper-modality to obtain non-verbal behavioral information. Then we use non-verbal behavior information as an auxiliary role to shift word representations. Since the CHFN dynamically shifted each word representation based on non-verbal cues across the utterance, multimodal data sources need not be pre-aligned by manual proposes or follow the same sampling rate (as required by previous works [10, 12, 25, 32]). Also, this fusion strategy can efficiently consider long-range dependencies between elements across modalities. Finally, the sequence of shifted word representations is fed into the multimodal transformer layers to predict the sentiments or emotions expressed in the utterance.

We evaluate our proposed model on two public and popular benchmark datasets: CMU-MOSI [39] and CMU-MOSEI [40]. The experiments show that the proposed method can produce better performance than the state-of-the-art method. In addition, ablation studies and further analysis show the effectiveness of our architecture.

The contributions of this paper are summarized as:

- We design a new multimodal interaction method dominant by the language modality. This method explores the underlying dynamics of multimodal language data by constructing multimodal interactions, and all modalities can be interacted with at once.
- We propose an efficient framework named Cross Hyper-modality Fusion Network (CHFN) to dynamically adjust word representations based on non-verbal cues across the utterance. In CHFN, modalities do not require pre-alignment operations, nor do they need to follow a similar sampling rate.
- We conduct comprehensive experiments on two publicly available datasets - CMU-MOSI and CMU-MOSEI. The proposed model outperforms the state-of-the-art (SOTA) aligned and unaligned methods.

2 RELATED WORKS

2.1 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) has become a hot topic in natural language processing [24, 25]. The expressive power of the single language modality is limited by the ambiguity of the language [33]. The ambiguity usually appears in scenarios, including slang, sarcasm, and so on. To overcome the limitation of the single language modality, the additional information from non-verbal behaviors can be a significant complement.

Most of the previous works in this area relied on the assumption that multimodal data sources are already aligned in the resolution of words. Zadeh et al. (2018a) used three separate LSTMs to model each modality. Then, it used Delta-memory attention and Multi-View Gated Memory to capture both temporal and inter-modal interactions. Wang et al. (2019) shifted word representations using non-verbal behavior information on aligned multimodal data. Moreover, long-range dependencies are not considered between elements across modalities. Hazarika et al. (2020) projected modality features into two different spaces (modality-invariant and modality-specific), and all information obtained from these spaces is concatenated together for the sentiment analysis. Although they have achieved excellent results in MSA tasks, these methods are limited to using aligned multimodal data sources. They need manually preprocess visual and acoustic features to align non-verbal behavior information with the resolution of the words. However, the manual alignment process is usually time-consuming and labor-intensive.

More general methods for the sentiment analysis task should be studied on unaligned multimodal data sources. Tsai et al. (2019a) extends the multimodal transformer architecture by using directional paired cross-attention to transform one modality into another. However, their proposed cross-modal interaction is a bi-modal operation that only accounts for the input from two modalities at a time. Therefore, they concatenated multiple bi-modal interactions to obtain multimodal representations, retaining the large number of parameters required for the original modal information. Yang et al. (2021) first convert unaligned multimodal sequence data into a graph. Then, an operation called MTAG is designed to capture the various interactions among modalities. In these methods, they consider the contribution of each modality is equal. However, recent works have concluded that textual features typically outperform non-textual features in the sentiment analysis task [19, 26, 27].

In our work, we proposed an end-to-end multimodal sentiment analysis method using unaligned multimodal data sources. It shifts the word representations based on non-verbal cues across the utterance by constructing the multimodal interactions between language and no-verbal behavior information. Furthermore, this interaction establishes the dominance of text features in multimodal sentiment analysis at the same time. To the best of our knowledge, there has been little research on modeling unaligned multimodal data sources with language modality as a dominant role.

2.2 Transformer and BERT

Transformer[31] is a sequence-to-sequence model without recurrent structure. It has been widely adopted in various fields, such as natural language processing (NLP), computer vision (CV) [1, 7, 22], and speech processing [3, 6, 9], especially for Pre-trained Models

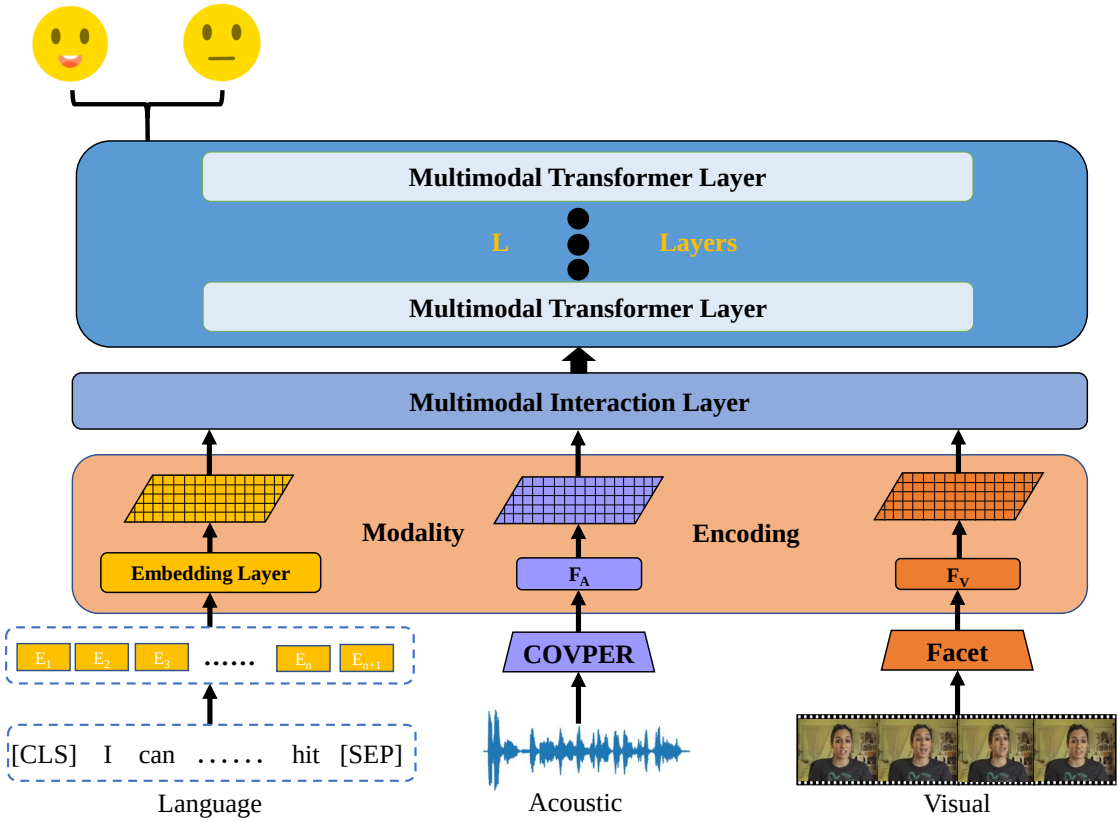

Figure 1: The overall architecture of our proposed Cross Hyper-modality Fusion Network. The model has three components: (1) modality encoding, (2) multimodal interaction layer, and (3) multimodal transformer layer. The raw multimodal signals are processed by the modality encoding to obtain numerical sequential vectors. Then, in the multimodal interaction layer, it generates the multimodal-shifted word representation by integrating the non-verbal shift vector to the original word embedding. Finally, the multimodal shifted word representations are fed into the multimodal transformer layer to predict the sentiment or emotion expressed in the sentences.

(PTMs) [15]. Recently, the pre-trained BERT (Bidirectional Encoder Representations from Transformers) [5], Transformer-based contextual word representations, has shown excellent performance in multiple disciplines within NLP [25]. The pre-trained BERT, which uses Transformer's encoder as the subject model, can capture bidirectional context to provide better representations than non-contextual PTMs. Fine-tuning of pre-trained BERT can bring very significant effect improvement for downstream tasks in the NLP. Therefore, the pre-trained BERT is chosen as the basis for sentiment analysis tasks in our work. Currently, the works utilizing the pre-trained BERT can be divided into two categories in MSA [35]. The first is to use the pre-trained BERT as the extraction module[12, 26, 35] for language modality features. The second is to integrate non-verbal behaviors information on the middle layers [25, 30]. In this paper, we use the second way. The pre-trained BERT is fine-tuned by using non-verbal behavior information to shift the word representations on the middle layers.

3 METHODOLOGY

This section introduces the task setup of multimodal sentiment analysis and the overall architecture of our proposed model. For our model, the signal processing of raw modality is first described in detail. Then, we introduce the method for integrating visual and acoustic information into words representations. Finally, we describe the training process of multimodal-shifted word representations, which integrate non-verbal behavior information.

3.1 Task Setup

The task of Multimodal Sentiment Analysis aims to judge the sentiment intensity or the emotion label using multimodal signals that we obtained from the datasets. These multimodal signals include spoken words (X_l), visual (X_v) and acoustic (X_a). The inputs of our model are unimodal raw sequences $X_m \in \mathbb{R}^{T_m \times d_m}$, where $m \in \{l, v, a\}$, T_m is the sequence length and d_m is the representation vector dimension of modality m. The model outputs $\hat{y} \in R$ as the final predictive result to fit the task.

3.2 Overall Architecture

The overview of our model is shown in Figure 1. It consists of three major parts: (1) *modality encoding* utilizes feature extractors (firmware for non-verbal behavior information such as visual and acoustic) and tokenizer (for language modality) to convert raw input multimodal signals into numerical sequential vectors (word, visual, and acoustic embedding). (2) *multimodal interaction layer* takes the original word embedding as well as visual and acoustic embedding as input. This layer will obtain multimodal-shifted word representations by integrating non-verbal behavior information to the original word embedding. (3) *multimodal transformer layer* trains multimodal-shifted word representations obtained from the previous layer and outputs a final label into judge the sentiment state. The following subsections discuss the details of our model.

3.3 Modality Encoding

For language modality, we use the Transformer-based pre-trained, BERT [5] as the text encoder. The raw sentence $s = \{w_1, w_2, \ldots, w_n\}$ is concatenated with two special tokens ([CLS] at the head and [SEP] at the tail) to form $s' = \{[CLS], e_1, e_2, \ldots, e_n, [SEP]\}$. Then, we input s' to the embedding layer of BERT which outputs numerical sequential vectors $X_l = \{l_0, l_1, \ldots, l_{n+1}\}$. For visual and acoustic modality, following previous works [11, 12], we use single-directional Long-Short-Term Memory (sLSTM) networks [13] to capture the temporal features of these modalities.

$$\begin{aligned} F_a &= \text{sLSTM}\left(I_a; \theta_a^{lstm}\right) \in R^{d_a} \\ F_v &= \text{sLSTM}\left(I_v; \theta_v^{lstm}\right) \in R^{d_v} \end{aligned} \tag{1}$$

The features of different modalities are enforced to have the same dimension by the convolutional fusion block controlling the kernel size of the 2D temporal convolutional operation used for visual and acoustic modalities.

$$X_{\{v,a\}} = \text{Conv 2D}\left(\{v, a\}, k_{\{v,a\}}\right) \in \mathbb{R}^{T_{\{v,a\}} \times d} \tag{2}$$

where $k\{v, a\}$ are the sizes of the convolutional kernels for the modalities $\{v, a\}$, and d is a common dimension. Convolved operations are important since the non-verbal behaviors are collected at different sampling rates. Moreover, the obtained features with the same dimensions facilitate the processing of subsequent behavior information. These outputs $\{X_l, X_a, X_v\}$ serve as the initial inputs to the multimodal interaction layer.

3.4 Multimodal Interaction layer

In multimodal data sources, the appearance of language modality is usually accompanied by non-verbal behavioral information consisting of visual and audio - simply facial expressions and intonations co-occurring with language. In different non-verbal contexts, the speaker's expressed intentions often vary dynamically. Non-verbal behavior information can impact the meaning of words, and therefore language and non-verbal behavior information jointly determine the position of words in the semantic space.

In our work, we use non-verbal behavior information as an auxiliary role to shift word representations. Figure 2 shows this process. The visual and acoustic modalities are concatenated together to obtain hyper-modality, that is, non-verbal behavioral information $X_\beta \in \mathbb{R}^{T_{(v+a)} \times T_d}$. We consider two sequences data (language X_l and non-verbal behaviors information X_β) from each of them denoted $X_l \in \mathbb{R}^{T_l \times d_l}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$, respectively. Similar to the self-attention mechanism in Transformers, the operation of integrating non-verbal behavior information into language modality also involves the query, key, and value, which are represented as $Q_l = X_l W_{Q_l}$, $K_\beta = X_\beta W_{K_\beta}$, and $V_\beta = X_\beta W_{V_\beta}$, respectively. The weights $W_{Q_l} \in \mathbb{R}^{d_l \times d_k}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$, and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$ are learnable parameters.

The attention coefficients a between language and non-verbal behaviors information is calculated as follows:

$$a = \frac{Q_l K_\beta^T}{\sqrt{d_k}} = \frac{X_l W_{Q_l} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}} \in \mathbb{R}^{T_l \times T_\beta} \tag{3}$$

To make the coefficients easily comparable, we normalize the attention coefficients using the softmax function:

$$e = \text{softmax}(a) = \text{softmax}(\frac{X_l W_{Q_l} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}) \tag{4}$$

The non-verbal shift information is calculated as follows:

$$\begin{aligned} H = eV_\beta &= \text{softmax}(a)V_\beta \\ &= \text{softmax}\left(\frac{Q_l K_\beta^\top}{\sqrt{d_k}}\right) V_\beta \\ &= \text{softmax}\left(\frac{X_l W_{Q_l} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta} \end{aligned} \tag{5}$$

where $H \in \mathbb{R}^{T_l \times d_k}$. These operations highlight the relevant information in visual and acoustic modalities conditioned on language modality. Note that the non-verbal shift H has the same length as X_l, but is meanwhile represented in the feature space of V_β.

Our proposed model learns to dynamically shift the word representations by integrating the non-verbal shift H into the original word embedding. And, we apply a scaling factor α to constrain the magnitude of the non-verbal shift H to be within a desirable range.

$$\bar{X_l} = X_l + \alpha H \tag{6}$$

$$\alpha = \min\left(\frac{\|X_l\|_2}{\|H\|_2}\lambda, 1\right) \tag{7}$$

where λ is a hyper-parameter selected through the cross-validation process. $\|X_l\|_2$ and $\|H\|_2$ donote the $L2$ norm of the X_l and H respectively. Finally, we apply a layer normalization and dropout layer to $\bar{X_l}$.

3.5 Multimodal Transformer Layer

BERT, which uses Transformer encoder as the subject mode, has shown excellent performance in multiple disciplines within NLP. It can provide better representation through capturing bi-directional context using Transformer. In this paper, we use the BERT variant (*BertForSequenceClassification*) to complete MSA tasks. The structure of the multimodal transformer layer is the same as the transformer encoder layer. We use the multimodal transformer layer to train the sequence of shifted word representations that integrate non-verbal behaviors information.

Given a n length language sequence $s = \{w_1, \ldots, w_n\}$, [CLS] and [SEP] tokens are appended to L so that our model can use [CLS] later for class label prediction. After modality encoding processing, we can obtain the original word embedding sequence $X_l = \{l_{CLS}, l_1, \ldots, l_{n+1}\}$. Then, in the multimodal interaction layer, it can obtain non-verbal behavior information related to the language modality by multimodal interaction. And, this related behavior information is integrated into the language modality to change the position of words in the semantic space. These shifted words representations can be denoted as $\bar{X}_l = \{l_{CLS}, l'_1, l'_2, \cdots, l'_{n+1}\}$. $\bar{X}_l$ is fed into the multimodal transformer layers that follow. $[CLS]$ token has the information necessary to make a class label prediction, and therefore the output l_{CLS} of the last multimodal transformer layer is used as a label for multimodal sentiment analysis.

4 EXPERIMENTS

This section introduces our experimental settings, including the experimental datasets, evaluation criteria, preprocessing, and baselines.

4.1 Datasets

The proposed algorithm is tested using two public benchmark multimodal sentiment analysis and emotion recognition datasets: CMU-MOSI [39] and CMU-MOSEI [40]. These datasets provide unaligned multimodal signals (language, visual, and acoustic) for each utterance. Here, we give a brief introduction to the above datasets.

CMU-MOSI: The CMU-MOSI is a commonly used dataset for human multimodal sentiment analysis. It consists of 2,198 short monologue video clips (each clip lasts for the duration of one sentence) expressing the opinion of the speaker inside the video on a topic such as movies. The utterances are manually annotated with a continuous opinion score between [−3, +3], [−3: highly negative, −2 negative, −1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, +3 highly positive].

CMU-MOSEI: The CMU-MOSEI is an improved version of CMU-MOSI. It contains 23,453 annotated video clips (about 10 times more than CMU-MOSI) from 5,000 videos, 1,000 different speakers, and 250 different topics. The number of discourses, samples, speakers, and topics is also larger than CMU-MOSI. The range of labels taken for each discourse is consistent with CMU-MOSI.

4.2 Preprocessing

We utilize the standard low-level features that are provided by the respective benchmarks.

Language Modality: Traditionally, language modality features have been GloVe [21] embeddings for each token in the utterance. GloVe features are 300 dimension token embeddings. However, recent works [12, 25] have demonstrated that BERT can provide better performance than GloVe in feature extraction. Therefore, BERT is used to obtain the features of the language modality in the proposed method. We utilize the *bert-base-uncased* pre-trained models.

Visual Modality: CMU-MOSI and CMU-MOSEI use Facet to extract facial expression features, including facial action units and facial gestures based on a Facial Action Coding System (FACS) [8]. This process is repeated for each sampled frame within the utterance video sequence.

Acoustic Modality: COVAREP [4] is used to extract the following relevant features: fundamental frequency, quasi-open quotient, normalized amplitude quotient, glottal source parameters (H1H2, Rd, Rd conf), VUV, MDQ, the first 3 formants, PSP, HMPDM 0-24 and HM-PDD 0-12, spectral tilt/slope of wavelet responses(peak/slope), MCEP 0-24.

The multimodal signals in our experiments were unaligned. The multimodal data sequences need not pre-align all three modalities following the convention established in [2].

Metrics	MAE	Corr	Acc-2	F1-Score	Acc-7
(Word Aligned) CMU-MOSI Sentiment					
MFN†	0.965	0.632	77.4/-	77.3/-	34.1
RAVEN†	0.915	0.691	78.0/-	76.6/-	33.2
MulT	0.871	0.698	-/83.0	-/82.8	40.0
MFM †	0.877	0.706	-/81.7	-/81.6	35.4
ICCN†	0.860	0.710	-/83.0	-/83.0	39.0
MISA	0.783	0.761	81.8/83.4	81.7/83.6	42.3
MAG*	0.727	0.781	82.37/84.43	82.50/84.61	43.62
(Unaligned) CMU-MOSI Sentiment					
TFN†	0.901	0.698	-/80.8	-/80.7	34.9
MulT	0.889	0.686	-/81.1	-/81.0	39.1
MTAG	0.866	0.722	-/82.3	-/82.1	38.9
Self-MM*	0.712	0.795	82.54/84.77	82.68/84.91	45.79
MMIM	0.700	0.800	84.14/86.06	84.00/85.98	46.65
CHFN	**0.689**	**0.809**	**84.3/86.4**	**84.2/86.2**	**48.6**

Table 1: Performances of multimodal models on CMU-MOSI. Best results are highlighted in bold. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive". NOTE: - means the result is not given in the paper; † is from [12]; * is from [11].

4.3 Evaluation Criteria

The evaluation metrics of previous works[12, 35] are referred to in the experiments. There are five evaluation metrics, namely Mean Absolute Error (MAE), Pearson Correlation (Corr), Binary Accuracy (Acc-2), F1-Score, and Seven-class Accuracy (Acc-7). MAE and Corr are regression tasks. Acc-2, F1-Score, and Acc-7 are classification tasks. For the calculation of Acc-2, two different evaluation methods are included. The first one is negative/non-negative classification [38], where non-negative includes neutral sentiment information. The second one is negative/positive classification [30], excluding neutral sentiment information. For MAE, lower values indicate better performance. For other metrics, higher values indicate better performance.

4.4 Baselines

The various state-of-the-art models introduced below are used as references for comparison.

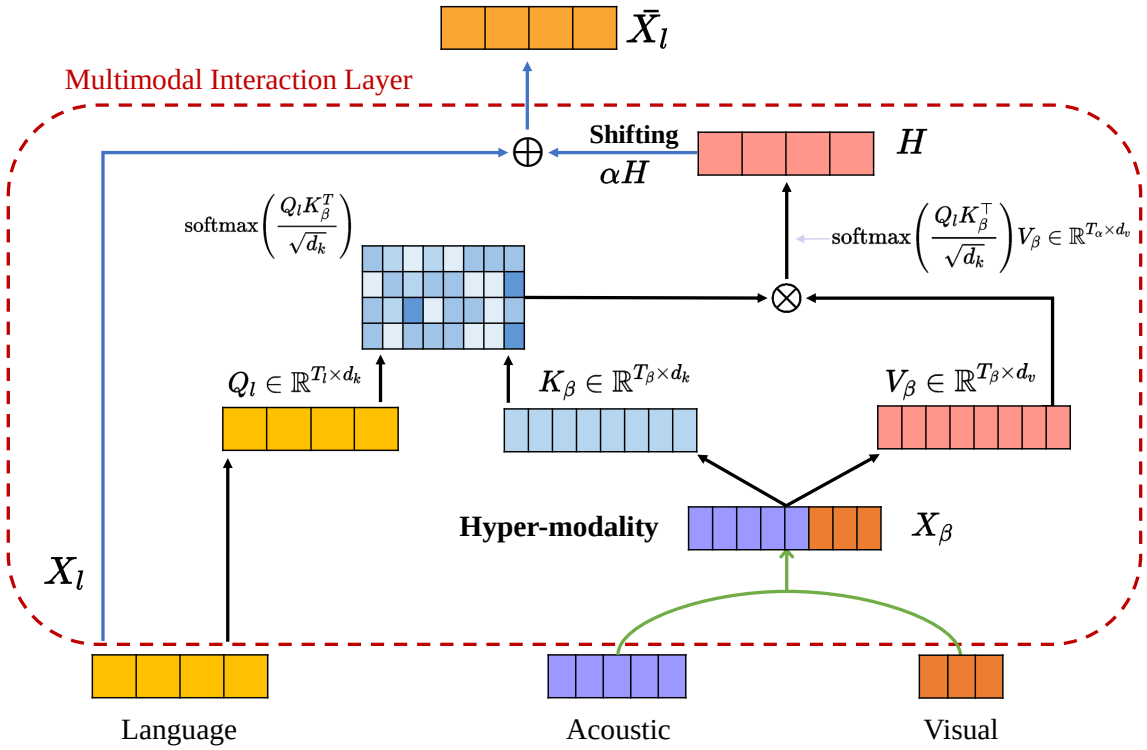

Figure 2: The details of the multimodal interaction layer. The layer takes as input word embedding, as well as visual and acoustic information. Firstly, the visual X_v and acoustic X_a are concatenated to form non-verbal behavior information X_β. Then we utilize the cross-multimodal attention mechanism to get the non-verbal behavior information H associated with the language modality. Finally, the behavior information H is used to shift the position of words in the semantic space.

TFN: Tensor Fusion Network (TFN) [36] performs an outer product of the output vectors after encoding the three modes to learn the intra- and inter-modal dynamics in an end-to-end manner and can capture uni-, bi-, and tri-modal interactions.

MFN: Memory Fusion Network (MFN) [37] uses three separate LSTMs to model each modality and uses Delta-memory attention and Multi-View Gated Memory to capture both temporal and inter-modal interactions.

MulT: Multimodal Transformer for Unaligned Multimodal Language Sequence (MulT) [30] extends the multimodal transformer architecture by using directional paired cross-attention to transform one modality into another.

RAVEN: The Recurrent Attended Variation Embedding Network (RAVEN) [32] utilizes a Gated Modality-mixing Network re-adjusting word embeddings according to auxiliary non-verbal signals on pre-aligned multimodal datasets.

MFM: Learning Factorized Multimodal Representations (MFM) [29] introduces a model that factorizes representations into two sets of independent factors: multimodal discriminative and modality-specific generative factors. The former is used for classification, and the latter is used to learn the modality-specific generative features.

ICCN: For Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis, Interaction Canonical Correlation Network (ICCN) [26] first extracts features from audio and video modalities, and then fuses them with text embeddings to get two outer products, text-audio, and text-video. Finally, the external products are fed into CCA network, and their output is used to predict.

MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (MISA) [12] combines various losses, including distribution similarity, orthogonal loss, reconstruction loss, and task prediction loss, to learn modality-invariant and modality-specific representations.

MAG: Integrating Multimodal Information in Large Pretrained Transformers (MAG-BERT) [25] is an improved work on RAVEN, which applies Multimodal Adaptive Gate (MAG) on different layers of the BERT backbone.

MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences [34] first convert unaligned multimodal sequence data into a graph. Then, an operation called MTAG is designed to capture the various interactions among multimodalities.

Self-MM: Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis

Metrics	MAE	Corr	Acc-2	F1-Score	Acc-7
(Word Aligned) CMU-MOSEI Sentiment					
MFN†	-	-	76.0/-	76.0/-	-
RAVEN†	0.614	0.662	79.1/-	79.5/-	50.0
MulT	0.580	0.703	-/82.5	-/82.3	51.8
MFM†	0.568	0.717	-/84.4	-/84.3	51.3
ICCN†	0.565	0.713	-/84.2	-/84.2	51.6
MISA	0.555	0.756	83.6/85.5	83.8/85.3	52.2
MAG*	0.543	0.755	82.51/84.82	82.77/84.71	52.67
(Unaligned) CMU-MOSEI Sentiment					
TFN†	0.593	0.700	-/82.5	-/82.1	50.2
MulT	0.591	0.694	-/81.6	-/81.6	50.7
Self-MM*	0.529	0.767	82.68/84.96	82.95/84.93	53.46
MMIM	0.526	0.772	82.24/85.97	82.66/85.94	54.24
CHFN	**0.525**	**0.778**	**83.7/86.2**	**83.9/86.1**	**54.30**

Table 2: Performances of multimodal models on CMU-MOSEI. Best results are highlighted in bold. In Acc-2 and F1-Score, the left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive". NOTE: - means the result is not given in the paper; † is from [12]; * is from [11].

(Self-MM) [35] proposes a label generation module based on self-supervised learning to obtain labels for each modality, joint multimodal, and unimodal learning to learn consistency and difference, respectively.

MMIM: Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis (MMIM) [11] proposes a hierarchical MI maximization framework that occurs at the input level and fusion level to reduce the loss of valuable task-related information.

For the tasks of MSA, the MMIM [11] stands as the state-of-the-art (SOTA) model on both CMU-MOSI and CMU-MOSEI.

5 RESULTS AND ANALYSIS

In this section, we discuss experimental results, ablation studies, and further analysis.

5.1 Performance Comparison

Table 1-2 show the results of our model in comparison with baseline models on the CMU-MOSI and CMU-MOSEI datasets. According to the requirement of data alignment, we divide the baseline models into two categories: Unaligned and Word Aligned.

As shown in both tables, the proposed model produces competitive results compared to the adopted baselines, which include the methods considering unaligned/aligned multimodal sequences. Compared to the methods that study multimodal fusion on manually pre-aligned datasets, our model can infer the long-range dependencies through dynamically shifting each word representation based on non-verbal cues across the utterance. Therefore, using the unaligned asynchronous multimodal sequences, our model still outperforms the models that utilize the word aligned sequences. Compared to the models that study multimodal fusion on unaligned

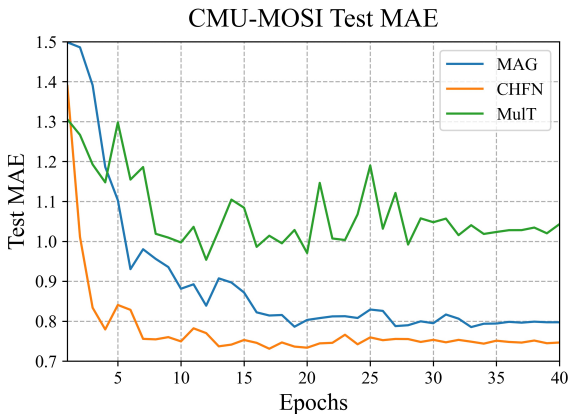

Figure 3: Convergence comparison of CHFN, MulT, and MAG on test set of CMU-MOSI dataset.

datasets, the proposed model can provide better performance. It should be noted that the results of the proposed model are better than MulT and MMIN that consider each modality is equal in the contribution to multimodal sentiment analysis. In our proposed model, we set the language modality as the dominant role, while the non-verbal behavior information serves as auxiliary roles to shift the word representations, which recent works have validated. Hence the results demonstrate that if three modalities are equal, this may impair the performance of multimodal representations. Furthermore, since the proposed model can process multiple modality signals at once, the results prove that more rich interactions can improve sentiment prediction.

To better illustrate how our proposed model outperforms the MAG and MulT, we compute the Mean Absolute Errors (MAE) of all the predictions on the test set of CMU-MOSI and paint their distributions in Figure 3. It can be seen that the curve of CHFN is lower than that of the MulT and MAG, which indicates that our model can achieve higher accuracy. These results preliminarily prove the effectiveness of our method in MSA tasks.

5.2 Ablation Studies

In our proposed method, the multimodal interaction layer is a core component. The component is explicitly designed to model subtle structures in verbal and non-verbal behavioral information. And it also integrates dynamic variations to the original word embeddings. Several ablation studies were performed to evaluate the impact of the component and to demonstrate the necessity of multimodal interaction within the component. The different versions of the model are explained as follows:

T: We remove the multimodal interaction layer in our proposed model, which causes the model to lose the ability to process behavioral information. Moreover, only the language modality is used for the sentiment analysis task. The original word representations aren't shifted by the non-verbal behavioral information that accompanies it.

T+A: We construct a bi-modal interaction between text and acoustic in the multimodal interaction layer. The audio information

Metrics	MAE	Corr	Acc-2
T	0.728	0.789	85.2
$T + A$	0.719	0.792	85.6
$T + V$	0.717	0.797	85.5
$T + A + V$	**0.689**	**0.809**	**86.4**

Table 3: Ablation studies on CMU-MOSI dataset. Best results are highlighted in bold.

that accompanies the language modality is used to shift the original word embeddings.

T+V: We construct a bi-modal interaction between text and visuals in the multimodal interaction layer. Visual information that accompanies the language modality is used to shift the original word embeddings.

T+V+A: We construct the multimodal interaction, including language, visual and audio, in the multimodal interaction layer at one time. The visual and acoustic information that accompanies the language modality are used to shift the original word embeddings.

Table 3 shows the results of the ablation studies performed on our model. The results show that both the multimodal interaction layer and achieving all modalities interactions at a time are necessary for achieving state-of-the-art performance. This further proves that our proposed adjustment of word representations with unaligned behavioral information is successful in the multimodal interaction layer. Another observation is that we have found that adding a modality can continuously bring performance improvements to our model. Overall, ablation studies show that dynamically shifting word representations in different non-verbal contexts is successful. And multimodal interactions are more robust than single modal or bi-modal for sentiment analysis tasks.

5.3 Further Analysis

In the field of multimodal sentiment analysis, the interactions between language modality and non-verbal behavior information change the perception of the language expressed sentiment. Our proposed CHFN can dynamically build text and non-verbal behaviors information interaction at the scale of the entire utterances. In order to show the learning capabilities of our model, we visualized the interactions information between language modality and non-verbal behavior information. Figure 4 displays the visualization of the multimodal interactions between elements across modalities. We can see that the CHFN can correlate the emotion-related textual word with the corresponding video frames and audio segments well in the long-range dependencies. This means that our model successfully captures the multimodal interactions between language modality and non-verbal behavior information. Note that we study the multimodal interactions in unaligned multimodal sequences data. And this interaction occurs at the scale of the entire utterances.

6 CONCLUSION

In this paper, we presented the end-to-end sentiment analysis method named Cross Hyper-modality Fusion Network (CHFN).

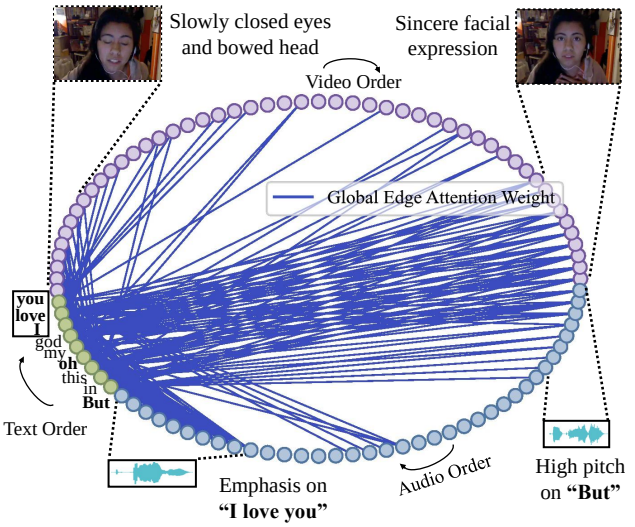

Figure 4: Example visualization of language and behavior information interactions. Each circle represents a node from video/text/audio modalities, and the blue lines denote significant interactions among modalities. We observe that there is a close relationship between language and non-verbal behavior information. The intensity of sentiment expressed by language varies in different non-verbal contexts. Such as "I love you" and the sincere facial, which indicates strong positive sentiment.

CHFN can model the multimodal interactions on unaligned multimodal data sources and builds multimodal-shifted word representations that dynamically capture the variations in different non-verbal contexts. Our model need not manually pre-process visual and acoustic modalities to align non-verbal behaviors information with the resolution of the words. At the same time, it can also solve the long-range dependencies of multimodal sequence data. CHFN achieves SOTA performance on two publicly available datasets, CMU-MOSI and CMU-MOSEI. Furthermore, we explored the impact of different non-verbal modalities on model performance through the ablation study. Finally, we also visualize the interaction information in the multimodal sentiment analysis task. The experimental results over different benchmarks clearly demonstrate that our proposed approach obtains better results than the existing state-of-the-art works. For multimodal sentiment analysis tasks, our model can capture the meaningful dynamic changes of textual information in non-verbal contexts.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.

[2] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 163–171.

[3] Xie Chen, Yu Wu, Zhenghao Wang, Shujie Liu, and Jinyu Li. 2021. Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5904–5908.

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5884–5888.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[8] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[9] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).

[10] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 6–15.

[11] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9180–9192. https://aclanthology.org/2021.emnlp-main.723

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

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. 2021. Attention Is Not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8148–8156.

[15] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A Survey of Transformers. *arXiv preprint arXiv:2106.04554* (2021).

[16] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2554–2562.

[17] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 481–492.

[18] Sijie Mai, Songlong Xing, Jiaxuan He, Ying Zeng, and Haifeng Hu. 2020. Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion. *arXiv preprint arXiv:2011.13572* (2020).

[19] Sijie Mai, Songlong Xing, and Haifeng Hu. 2021. Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1424–1437.

[20] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

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

[22] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

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

[24] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing* (2020).

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

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

[27] Zhongkai Sun, Prathusha K Sarma, Yingyu Liang, and William Sethares. 2021. A New View of Multi-modal Language Analysis: Audio and Video Features as Text "Styles". In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1956–1965.

[28] Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong. 2021. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5301–5311.

[29] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In *ICLR*.

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

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

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

[33] Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing Emotions in Video Using Multimodal DNN Feature Fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Association for Computational Linguistics, Melbourne, Australia, 11–19. https://doi.org/10.18653/v1/W18-3302

[34] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 1009–1021. https://www.aclweb.org/anthology/2021.naacl-main.79

[35] Wenmeng Yu, Hua Xu, Yuan Ziqi, and Wu Jiele. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[36] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.

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

[38] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

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