ITC 1/53
Information Technology and Control
Vol. 53 / No. 1 / 2024
pp.206-219
DOI 10.5755/j01.itc.53.1.35060

Tri-CLT: Learning Tri-Modal Representations with Contrastive Learning and Transformer for Multimodal Sentiment Recognition

Received 2023/09/09 | Accepted after revision 2024/01/24

HOW TO CITE: Yang, Z., Li, Z., Zhu, D., Zhou, Y. (2024). Tri-CLT: Learning Tri-Modal Representations with Contrastive Learning and Transformer for Multimodal Sentiment Recognition. *Information Technology and Control*, 53(1), 206-219. https://doi.org/10.5755/j01.itc.53.1.35060

Tri-CLT: Learning Tri-Modal Representations with Contrastive Learning and Transformer for Multimodal Sentiment Recognition

Zhiyong Yang
The College of Big Data and Internet of Things, Chongqing Vocational Institute of Engineering, and the College of Computer and Information Science, Chongqing Normal University, Chongqing, 402246, China;
e-mail: zyy@cqvie.edu.cn

Zijian Li, Dongdong Zhu
School of Computer and Information Science, Chongqing Normal University, Chongqing, 401331, China

Yu Zhou
The College of Finance and Tourism, Chongqing Vocational Institute of Engineering, Chongqing, 402246, China

Corresponding author: zyy@cqvie.edu.cn

Multimodal Sentiment Analysis (MSA) has become an essential area of research to achieve more accurate sentiment analysis by integrating multiple perceptual modalities such as text, vision, and audio. However, most previous studies failed to align the various modalities well and ignored the differences in semantic information, leading to inefficient fusion between modalities and generating redundant information. In order to solve the above problems, this paper proposes a transformer-based network model, Tri-CLT. Specifically, this paper designs Integrating Fusion Block to fuse modal features to enhance their semantic information and mitigate the secondary complexity of paired sequences in the transformer. Meanwhile, the cross-modal attention mechanism is utilized for complementary learning between modalities to enhance the model performance. In addition, contrastive learning is introduced to improve the model's representation of learning ability. Finally, this paper conducts experiments on CMU-MOSEI aligned and unaligned data, and the experimental results show that the proposed method outperforms the existing methods.

KEYWORDS: Multimodal Sentiment Analysis, fusion, transformer, cross-modal attention, contrastive learning.

1. Introduction

The thriving development of digital communication, online platforms, and social media has given people increasingly diverse ways to express their emotions and opinions. In addition to traditional text, there are now various modalities of data, including audio and video. The proliferation of this multimodal data has spurred the rise of multimodal sentiment analysis [25]. Multimodal Sentiment Analysis (MSA) is an emerging field involving sentiment recognition techniques for multiple perceptual modalities (e.g., text, vision, and audio). Compared with unimodal sentiment recognition, multimodal sentiment integrates multiple perceptual modalities to understand people's emotional states more comprehensively and accurately. In addition, multimodal sentiment recognition is essential in application scenarios such as intelligent customer service, social robotics, mental health assistance, and online education [1]. Simulating the way of communication in human life helps to build a more natural and convincing human-computer interaction experience.

In multimodal sentiment analysis tasks, effectively fusing features from visual (often contained in images or videos), audio (in the form of sound waves), and textual information (in the form of words) is crucial to improve performance. In existing fusion methods, simply splicing features from different modalities may not be able to fully capture inter-modal correlation information, such as Early Fusion [5, 23] and Late Fusion [36]. In contrast, Intermediate Fusion [5] can capture a certain degree of cross-modal relationships at the intermediate level and retain a certain degree of modality-specific information. Meanwhile, Tensor-based fusion methods [20] are getting more and more attention due to their high expressive ability across modalities. GAN [24] based methods can represent the correlation between multimodal data well and strengthen the inter-modal correlation information, but the computational resource requirement is high. In addition, Graph [21, 34] based methods use graph fusion networks to fuse modalities, and fusion methods based on Attention Mechanism [13, 22] and Deep Neural Network Fusion [16] require the design of specific structures for different tasks, which is more complex to implement. Cross-modal attention [15, 22] is a practical approach for integrating textual, visual, and audio semantic features, which takes advantage of the complementarity between modalities by using one modality to learn the contextual information of the other, thus obtaining a more expressive representation of the fused features. However, past approaches have often failed to align multimodal data effectively and ignored the differences in semantic information between modalities. According to recent findings [12], text is more important than audio and visual and contains advanced semantic information.

The essence of multimodal fusion is to integrate the features of different modalities into a unified embedding space, project the inputs into a shared embedding space, and capture the information between different modalities by representing different modalities in the common embedding space. In recent years, transformers have achieved excellent results in the multimodal domain [31-32]. People have proposed using transformers to learn the embedding space, using multiple independent transformers [3, 18], or using one transformer to learn the embedding space on video data [2]. Contrastive learning has shown strong performance in representation learning [9, 14]. It projects positive and negative samples into the embedding space to learn the encoding between modalities through contrastive loss, which helps to improve multimodal representation. To effectively learn the multimodal embedding space, this paper projects the inputs into a shared embedding space, uses a transformer as an intermediate processing step, and drives co-occurring modal embeddings closer together in the shared space utilizing contrastive learning, thus reducing the risk of overfitting and improving the model's representational learning ability.

The paper proposes an effective fusion network model for learning tri-modal representations with contrastive learning and Transformer (Tri-CLT) for Multimodal Sentiment Analysis. Figure 1 presents an overall of the model. Specifically, the model presented in this paper takes raw visual, audio, and textual data as inputs and extracts high-level features from the raw input data using a network architecture specific to each modality. The transformer fuses the multimodal information, and we design the Integrating Fusion Block to restrict the information interactions between modalities to the module, thus alleviating

the secondary complexity of paired sequences in the transformer. The resulting fused features enhance the semantic information of the three modalities. Next, cross-modal attention is employed to enhance paired feature information, achieving information reinforcement and complementarity to improve the performance of sentiment recognition. Meanwhile, we leverage contrastive learning to integrate embedding spaces from different modalities and map inputs with semantic similarity into each other's compact representations, thereby improving performance. Unlike the transformer model's existing architecture that uses the [cls] tokens' output for classification, Tri-CLT uses averaging the entire output sequence for sentiment prediction. Finally, a linear layer summarizes the fused feature information to obtain the sentiment prediction results. In this paper, we evaluate Tri-CLT on CMU-MOSEI [37]. The main contributions of this thesis can be summarized as follows:

This paper proposes the Tri-CLT model based on the transformer, which takes the combined data of the three modalities as input, fully utilizes the similarity and complementarity among modalities, and effectively fuses the interaction information of the three modalities. The model learns fused representations of multimodal sentiment recognition from CMU-MOSEI aligned and unaligned data. Experimental results show that Tri-CLT outperforms existing methods.

An Integrating Fusion Block module is designed to compensate for the difference in semantic information between different modalities, slow down the secondary complexity of paired modalities in the transformer, reduce redundant information, and combine with cross-modal attention to handle cross-modal related information to achieve fair and effective complementary learning.

Cross-modal contrastive learning is utilized to integrate the embedding space of different modal information, using a transformer as an intermediate processing step to improve the learning ability of multimodal embedding space.

2. Related Work

This section will divide the work into Multimodal Transformers and Multimodal Sentiment Analysis. Each section will review the existing research work.

2.1. Multimodal Transformers

The remarkable success of transformer in natural language initially used in sequence-to-sequence machine translation tasks, VIT [7], and the proposed AST [10] have proven to be very effective in visual and audio modalities, making transformer [33] shine in different domains. Recently, Cheng et al. [6] designed a shared attention network for synchronizing audio and vision. Luo et al. [19] inherited their idea to learn the correspondence between audio and visual samples and proposed a transformer model that combines visual and text. Bain et al. [3] focused on temporal and spatial issues and processed modal information in two separate transformers. Recent studies use specialized transformer models applicable to different modalities and perform multimodal fusion by contrastive loss. Shvetsova et al. [27] proposed using a transformer encoder to represent three modalities and designed a contrastive loss considering multiple modal inputs. Akbari et al. [2] learned multimodal representations from raw data in a self-supervised environment. They designed a transformer-generalized multi-task architecture with combined visual-text and visual-audio for contrastive loss.

However, most of the work relies on complex fusion strategies that introduce more parameters and computational costs and are prone to overfitting. Nagrani et al. [22] proposed a Multimodal Bottleneck Transformer (MBT), which restricts the exchange of information between modalities during the fusion process. Inspired by their work, we will limit the information exchange between pairs of modal sequences within the transformer to avoid over-computation and reduce computational complexity.

2.2. Multimodal Sentiment Analysis

The continuous advancement of deep learning has made the processing of multimodal data more efficient, leading to widespread attention to multimodal sentiment analysis [4, 28]. Researchers in the past primarily embodied their work in designing complex multimodal fusion strategies. At early stages, Early Fusion [5, 23] is a method that directly connects the features of different modalities to form a comprehensive feature vector for training and decision making. Late Fusion [36] makes decisions independently on each modality. Then, it fuses the decision results of different modalities by weighted averaging. As fusion methods con-

tinue to deepen, tensor-based fusion is highly expressive regarding cross-modal dynamics. Zadeh et al. [35] designed the TFN to model the interaction between modes using tensor and perform geometric operations in feature space. Liu et al. [17] developed the LMF to learn cross-modal dynamics using low-rank decomposition. In recent research, Fu et al. [8] fuse visual and audio modalities by introducing NHFNet to enhance semantic features. Tsai et al. [31] proposed MulT, which performs cross-attention in paired modalities, for unaligned multimodal emotion recognition tasks. MISA presented by Hazarika [12] decomposes the modal features of the joint space, divides the modalities into specific and invariant, and maps the modalities to two subspaces. Rahman et al. [26] introduce MAG-BERT, which incorporates a fusion gate to learn the associations between the modalities using word boundary alignment, and Han et al. [11] use MMIM to maximize the multimodal information and apply the mutual knowledge on multimodal feature fusion, which in turn optimizes multimodal representations.

While most previous works have achieved some performance improvements in multimodal sentiment recognition, there still needs to be the problem of semantic differences between the three modalities and the presence of redundant information in the fusion process. This paper proposes a new fusion network that reduces the semantic gap between modalities and information redundancy in the fusion process while obtaining better multimodal representation through contrastive learning to achieve more accurate emotion prediction.

3. Methodology

The paper aims to design an efficient multimodal fusion architecture that integrates different modal features and projects different modalities into the multimodal embedding space to make their semantic inputs close to each other, thereby improving model performance and achieving accurate sentiment prediction. This section describes the proposed method in detail.

3.1. Standardized Sequence

Tri-CLT receives text, audio, and visual data from a given video. We use the same feature extraction method as the baseline [12], using specific networks for different modalities to extract features of the respective modalities, obtaining three sets of unimodal sequences $x_m \in R^{n\times l_m\times d_m}$. The standard batch size is denoted by n, and we use l_m to represent the length of the input sequence, and m represents text, visual, and modalities. d_m is the input dimension of the modal representation vector. We will individually pass the feature vectors extracted from each modality-specific backbone through the transformer encoder layer for feature extraction. Three sets of tokens will be generated: text $[t_{i1}, ..., t_{im}]$, visual $[v_{i1}, ..., v_{ik}]$, and audio $[a_{i1}, ..., a_{in}]$. In this process, we normalize the length of the sequence for each batch of inputs and convert the sequence into a standard input form that the transformer can accept as follows:

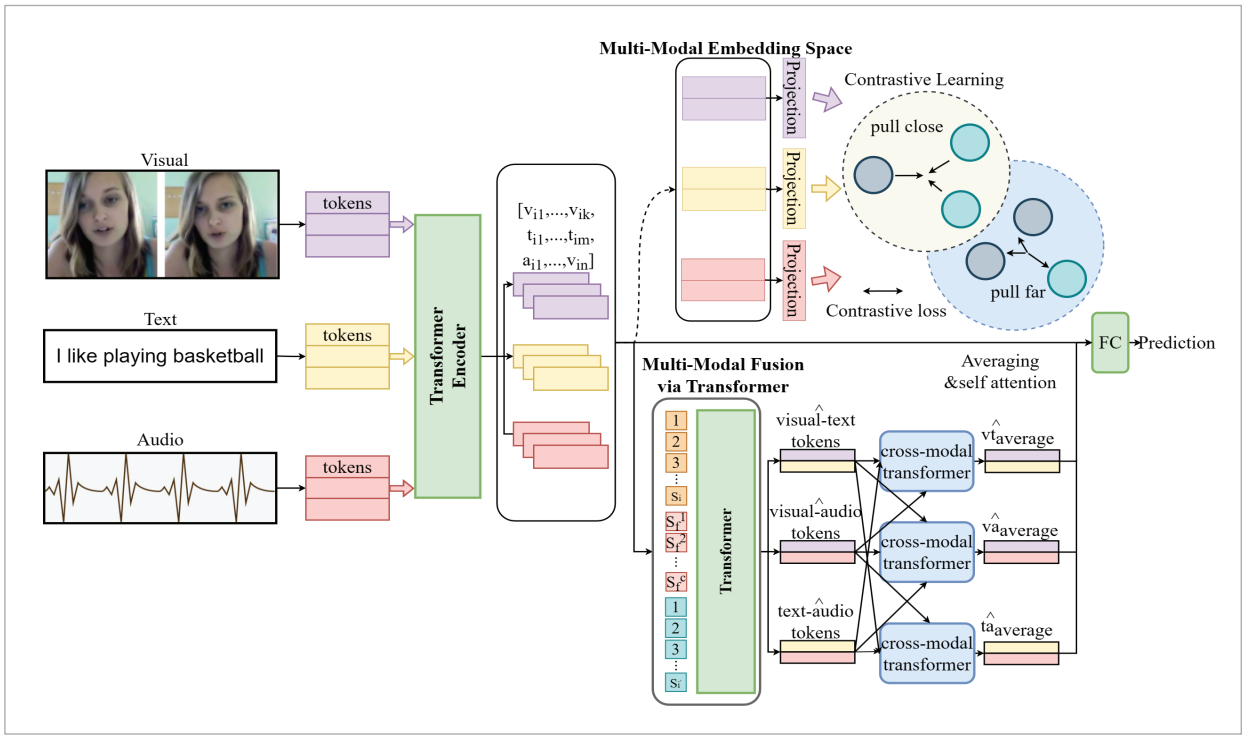
$$S_m=g(X_m;E,S_{cls})=\left[S_{cls},E_{x_m1},E_{x_m2},\ldots,E_{x_mN}\right]+P,\quad m\in\{\text{visual, audio, text}\}. \tag{1}$$

However, Tri-CLT differs from the traditional transformer processing method [12, 22]. It does not add any category embedding information and does not add positional embedding at a standardized sequence. The reasons are as follows: We divide CMU-MOSEI into aligned and unaligned data in our work. Using position embedding to process aligned data can help the model capture the order and position information in the input data. However, when dealing with unaligned data, inconsistent temporal patterns may add noise to the training process and thus negatively affect the model. Therefore, omitting positional embedding is a better choice, and it helps to handle variable-length sequences in the actual inference process. Additionally, we often use [cls] tokens to represent the overall picture of the data. Since the token information extracted by their respective backbones already has their own "fingerprints," there is no need to use [cls] tokens, which will prove beneficial to the model during ablation experiments.

3.2. Multi-Modal Fusion via Transformer

Effectively fusing information from multiple modalities is the goal of this paper. Since textual feature representation is more intuitive and visual and audio contain a substantial amount of redundant information, taking into account the equitable interaction of multiple modalities within the transformer, we propose a transformer-based fusion strategy, where the

Figure 1
The overall architecture of Tri-CLT



extracted three sets of modal tokens are input to the transformer for pairwise fusion. Usually, the computational complexity of the transformer grows linearly with the length of the sequence because the attention layer in the transformer touches and processes each token, thus generating a large amount of computational waste and affecting the model's performance. In order to solve the above problems, this study proposes a transformer-based feature fusion approach, as shown in Figure 1. Based on [22], we constrain cross-modal connections to intermediate layers and define the Integrating Fusion Block module, denoted as $S_f = S_f^1 S_f^2, ..., S_f$. Because S_f dimension is much smaller than the first two, the amount of computation from N^2 to N. This operation ensures that all cross-modal attention flows in the model share information only through these Integrating Fusion Blocks, thereby reducing redundancy and effectively overcoming the secondary complexity of the transformer to deal with the paired modal sequence.

In this study, we use only the encoder part of the transformer for fusion. In the transformer, an entire encoder layer comprises the following components: a Multi-Head self-attention (MSA), two LayerNorm (LN) transforms, and a Multilayer Perceptron (MLP), and these modules are connected sequentially through residual connections. Each layer representation is defined as S_m^{1+1} =Transformer(S_m^l). The process equation is as follows:

$$y^l = MSA\left(LN\left(S_m^l\right)\right) + S_m^l \quad m \in \{visual, audio, text\}. \quad (2)$$

$$S_m^{l+1} = MLP\left(LN\left(y^l\right)\right) + y^l \quad m \in \{visual, audio, text\}. \quad (3)$$

In the MSA layer, Querys and Keys perform dot product attention operations to obtain the similarity between feature vectors. Because of the nature of the multi-headed attention mechanism, Querys, Keys, and Values are obtained from the same vector S_m by different mappings. MSA(S_m) = Attention ($W^Q S_m$, $W^K S_m$, $W^V S_m$). The MLP layer captures features in the input sequence and the dependencies between sequences by introducing a nonlinear activation function.

We introduce the Integrating Fusion Block to fuse the modal features. In this paper, C (C is much smaller than l_m) Integrating Fusion Blocks are introduced in the input sequence of the transformer. The input sequence (in the case of vision and audio fusion) is now:

$$S=[S_v^1,\ldots,S_v^d||S_f^1,\ldots,S_f^c||S_a^1\ldots S_a^d], \quad (4)$$

where || represents the splice operation, we will update the Integrating Fusion Block between pairs of modalities twice. Each time, two modalities share information within the block, reducing redundancy and computational load and improving and maintaining model performance. For layer l, the computation process is as follows:

$$[S_i^{l+1}||\hat{S}_{f_i}^{l+1}]=\text{Transformer}\left([S_i^l||S_{f_i}^l];\theta_i\right). \quad (5)$$

This study considers three modal combinations: visual and text, text and audio, as well as visual and audio. The combined features of the visual and audio modalities are denoted as v_ia_i, the combined features of text and visual are represented as v_it_i; and the combined features of text and audio are represented as t_ia_i. During each training iteration, we apply the multimodal fusion Transformer three times to each sample i. To obtain the fused representation, we use a fused tokens list to represent, such as $v_it_i[v_{i1}, ...,v_{im} f_B^1, ..., f_B^c t_{i1}, ..., t_{in}]$ The multimodal input data are fused by the transformer to get the augmented data $\widehat{v_it_i}[\hat{v}_{i1}^{vt}, ..., \hat{v}_{im}^{vt}\hat{f}_B^1, ..., \hat{f}_B^c\hat{t}_{i1}^{vt}, ..., \hat{t}_{in}^{vt}]$ (where the superscript vt denotes attention to both modal visual and modal text). In this way, the modalities can learn from each other, making it possible to share information between the modalities and to fuse multimodal feature information effectively.

3.3. Multimodal Contrastive Learning

Even though the modal information is augmented with each other by the above methods, the semantic information is still very different. Contrastive loss can be used for representation learning, where the basic idea is that anchors and positive samples are continually brought closer together in the embedding space, and anchors and negative samples are pushed farther away. This process projects semantically similar inputs between modalities to positions close to each other. Specifically, we normalize the single-modal information [a, v, t] obtained through the transformer encoder layer and project it into the shared embedding space. We use this normalization process to align the magnitudes of the vectors, and only the angles between the vectors are considered when calculating the point similarity. For each anchor sample, its comparative loss formula is as follows:

$$L_{(m,\dot{m})}=-\log\left(\frac{\exp(m^T\dot{m}/\tau)}{\sum_{i=1}^B \exp(m_i^T\dot{m}_i/\tau)}\right), \quad (6)$$

where $L_{(m,\dot{m})}$, τ and B represent the contrastive loss between modality m and modality $\dot{m}$, temperature and batch size, respectively. In this work, we apply contrastive learning to these three modalities to enhance their interactions and increase the distinctiveness of fused representations among samples. Moreover, the contrastive loss L_{v_t}, L_{t_a} and L_{v_a} between the three unimodal information is obtained, L_{v_t} corresponds to (v, t), L_{t_a} corresponds to (t, a), and L_{v_a} corresponds to (v, a). Combining the above representations, the three loss results are combined and output:

$$L=\omega_{v\_t}L_{v\_t}+\omega_{t\_a}L_{t\_a}+\omega_{v\_a}L_{v\_a}, \quad (7)$$

where $\omega_{m\_\dot{m}}$ denotes the weighting factor of mode $(m, \dot{m})$.

3.4. Cross-modal Attention

The cross-modal attention mechanism can exploit the complementarity between modalities to reinforce each other by learning each other's feature information. To this end, we will define a cross-transformer layer that allows the information to interact through the attention layer, reinforcing the three sets of fused features with each other.

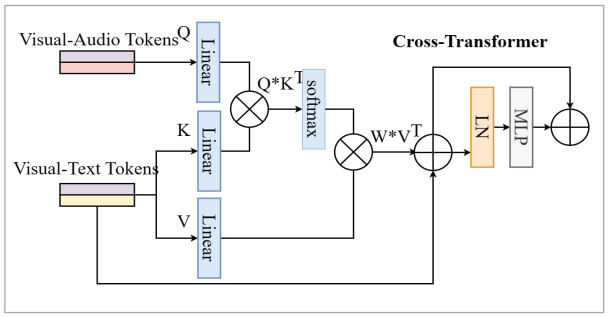
Figure 2 shows an overview of the cross-transformer. Here, we exemplify the mutual reinforcement between va and vt:

$$S_{va}^{l+1}=\text{Cross-Transformer}\left(S_{va}^l,S_{vt}^l;\theta_{va}\right). \quad (8)$$

$$S_{vt}^{l+1}=\text{Cross-Transformer}\left(S_{vt}^l,S_{va}^l;\theta_{vt}\right). \quad (9)$$

This paper defines tensor X and tensor Y for Multi-Headed Cross Attention (MCA), with X serving as the query, while the other tensor is employed as keys and values for weighted queries, defined as MCA(X, Y) = Attention (W^QX, W^KY, W^VY). The paired fusion fea-

Figure 2
Cross-Transformer



tures va and vt will follow the operations of the original transformer, but Equation (2) will become:

$$y^{l}=\mathrm{MCA}\left(\mathrm{LN}\left(S^{l}_{\{va,vt\}}\right),\mathrm{LN}\left(S^{l}_{\{vt,va\}}\right)\right)+S^{l}_{\{va,vt\}}\ . \quad (10)$$

3.5. Prediction

Eventually, we separately obtain the unimodal text features t, audio features a, visual features v, and paired fusion enhanced visual-text features $\hat{vt}^{va}$ and $\hat{vt}^{ta}$, visual-audio features $\hat{va}^{vt}$ and $\hat{va}^{ta}$, and text-audio features $\hat{ta}^{va}$ and $\hat{ta}^{vt}$. The unimodal features are enhanced through the self-attention mechanism, resulting in enhanced text feature $\hat{t}$, enhanced visual feature $\hat{v}$, and enhanced audio feature $\hat{a}$.These output features are averaged and summed $P=\hat{a}_{average}+\hat{v}_{average}+\hat{t}_{average}+\hat{vt}_{average}+\hat{va}_{average}+\hat{ta}_{average}$ ($\hat{vt}_{average}$ denotes taking the mean for the fusion features $\hat{vt}^{va}$ and $\hat{vt}^{ta}$). We feed the final information obtained into a linear layer to obtain the corresponding sentiment score.

4. Experiments

First, this section discusses the selection of datasets and the setup of experimental details. Second, Tri-CLT is compared with other recent work to evaluate its performance. Furthermore, to verify the effectiveness of each module and hyperparameters of Tri-CLT, we conducted ablation experiments for each task. Meanwhile, we discussed the setting of the hyperparameter ω. Finally, we visually analyze the multimodal representations to verify the method's effectiveness. A single RTX3080 is used to train the model in this work during the experiments.

4.1. Datasets

We have chosen to evaluate the model's performance on CMU-MOSEI in this paper. CMU-MOSEI is a mainstream dataset for Multimodal Sentiment Analysis developed by the MultiComp Lab at Carnegie Mellon University. CMU-MOSEI is a large human dataset that contains 23,453 sentences from YouTube videos on 1000 different speakers and 250 topics. The sentences were sourced from online video-sharing sites and included expressions of opinion on topics such as movies. There was a balanced gender distribution in the dataset (57% male, 43% female), and the average sentence length was 7.28 seconds. In the CMU-MOSEI dataset, a human artificially labeled each sample using an affective score ranging between -3 and 3. These scores represent different levels of emotion, from negative to positive. The dataset division includes training, validation, and test sets containing 16,322, 1,871, and 4,659 samples. According to recent studies, evaluation metrics include mean absolute error (MAE), Pearson's correlation coefficient (corr), F-score, and dichotomous and multiclassification accuracy. In dichotomization, (-3, 0) is considered negative, and (0, 3) is positive.

4.2.1. Backbone

This paper follows the previous baseline network setup [12] to ensure model comparability. For the visual modality, we use FaceNet to obtain facial sentiment features. For the audio modality, we extract acoustic features from COVAERP. These low-level statistical features include MFCC, pitch, voiced/unvoiced segmentation features, sound quality features, and other emotion-related features. For the text modality, pretrained BERT was used as the text feature extractor, and these backbones were fixed and not fine-tuned on the dataset. Finally, 768-dimensional features are obtained for text, 35-dimensional features for vision, and 74-dimensional features for audio.

4.2.2. Experimental Settings

This paper divided the CMU-MOSEI data into aligned and unaligned data. Tri-CLT will be experimented with both sets of data separately. Regarding the processing of aligned data, the data is processed using CMU MultimodalSDK v1.2.0 to obtain aligned data, and for the setup of unaligned data, the unaligned data is processed directly. Hyperparameters employed during the training and testing of CMU-MOSEI are presented in Table 1.

Table 1
Hyper Parameter Setting

Setting	CUM-MOSEI
Optimizer	Adam
Batch Size	16
Learning rate	1e-3
Attention head	8
Transformer Layer	4
Sequence length	50
Feature size	32

The model is optimized with the MSE loss L^{task} and the combined contrastive loss L^{cl}. A more considerable weight is set in Equation (7) for the corresponding loss of the text: $\omega_{t\_v} = \omega_{t\_a} = 1$, $\omega_{v\_a} = 0.1$, which is beneficial for training on CMU-MOSEI. We set the overall loss function as follows: $L^{Total} = L^{task} + \omega L^{cl}$, with $\omega = 0.7$.

4.3. Comparsion with Baselines

This section compares Tri-CLT with baselines using aligned and unaligned data on the CMU-MOSEI, respectively. We show quantitative results between Tri-CLT and other baselines in Tables 2-3. The results demonstrate that Tri-CLT performs better than previous methods for some metrics on the CMU-MOSEI dataset.

Table 2
Results on CMU-MOSEI under aligned data, comparison with baselines and existing methods

Model	F1	Corr	Acc-2	Acc-7	MAE
EF-LSTM	79.74	0.624	79.31	46.6	0.622
LF-LSTM	80.60	0.652	80.65	49.6	0.656
MFM[30]	84.31	0.703	84.42	51.3	0.568
MulT[31]	82.31	0.713	82.51	51.8	0.580
ICCN[29]	84.15	0.713	84.18	51.58	0.565
MFN[38]	80.63	0.670	79.60	49.1	0.618
MISA[12]	83.97	0.724	84.23	**52.20**	0.568
MAG-BERT[26]	84.71	**0.778**	85.21	51.9	**0.548**
Tri-CLT (Ours)	**85.33**	0.750	**85.55**	52.18	0.568

Tri-CLT outperforms all baselines in the aligned data with an F1 score of 85.33% and Acc-2 of 85.55%. In Table 2, we categorize existing methods as follows: 1) In earlier work, Early Fusion LSTM (EF-LSTM) fuses the input data features before inputting them into the LSTM network, and then the fused features are inputted into the LSTM for processing. Late Fusion LSTM (LF-LSTM) extracts each modal feature separately to make decision inference, and the different modal information of different modalities is fused through the voting mechanism. Tri-CLT improved by about 5% on F1 values and Acc-2 scores and by about 4.5% and 2.5% on Acc-7, respectively. 2) MFM [30] utilizes multimodal information and performs factor decomposition in the joint representation of different

Table 3
Results on CMU-MOSEI with unaligned data, compared to baselines and existing methods

Model	F1	Corr	Acc-2	Acc-7	MAE
TFN[35]	82.09	0.704	82.66	50.13	0.60
MulT[31]	81.9	0.699	81.48	50.6	0.591
LMF[17]	82.44	0.656	82.06	49.33	0.611
MMIM[11]	84.68	**0.743**	84.61	**53.11**	**0.552**
MISA[12]	81.1	-	81.7	52.1	-
Tri-CLT	**85.12**	0.741	**85.22**	51.94	0.560

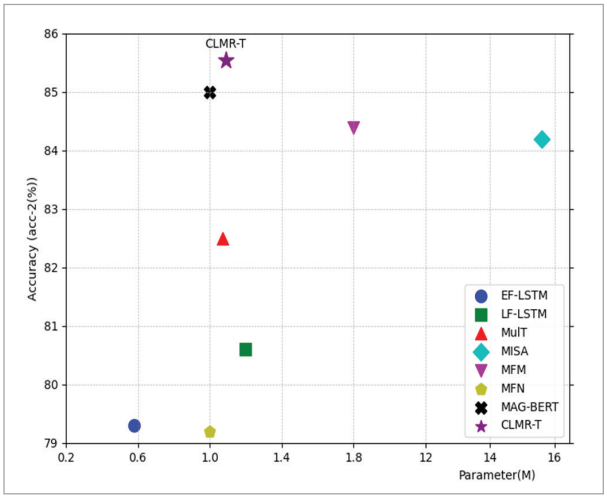
modalities. The correlation between modes is captured efficiently. Tri-CLT improves the F1 value, Acc-2, and Acc-7 scores by 1%, 1.1%, and 0.8%, respectively. 3) In a Cross-modal attention-based approach to multimodal fusion, MulT [31] uses a cross-modal transformer to learn multimodal representations, and ICCN [29] obtains two bimodal representations and feeds them into a network to generate trimodal representations and make predictions. Tri-CLT has significant improvements over these two architectures and achieves superior performance. 4) MFN [38], which uses delta-attention modules for interaction and summarization through multi-view gating networks. Compared with it, the F1 value, Acc-2 and Acc-7 scores of Tri-CLT were improved by 5%, 6%, and 3%, respectively. 5) MISA [12] is a method that focuses on modal invariance and specificity, which controls the representation space, while Tri-CLT focuses more on modal feature representation and achieves better results. 6) MAG-BERT [26], a method using multi-modal adaptive fusion gates, can achieve more than 85% on ACC-2, but Tri-CLT performs better than it.

In addition, this paper also tests on unaligned data. Some of the metrics of Tri-CLT outperform the baseline. among them, the F1 scores 85.12%, and the Acc-2 is 85.22%. We draw the following conclusions based on the results in Table 3. 1) Methods TFN [35] and LMF [17] created fused joint representations, and compared to these two methods, Tri-CLT improved the F1 value and Acc-2 by approximately 3%. 2) Compared with MISA and MulT in unaligned data, Tri-CLT can show better performance. 3) The method MMIM [11], which maximizes the interactive information between modes, can effectively control the modal representation space. Our method improves by about 0.6% on F1 and Acc-2.

Moreover, this paper also compares the efficiency of multiple models, and Tri-CLT shows excellent performance in smaller models. As shown in Figure 3, we compared the parameters of the fusion network and Acc-2, where the parameters of the unimodal feature extraction network are not included in the number of parameters. It demonstrates that the proposed fusion network on CMU-MOSEI can achieve higher accuracy with fewer parameters, achieving an excellent balance between parameters and performance.

Tri-CLT outperforms all other baselines in efficiency, which suggests that the proposed lightweight network is more suitable for real-world scenarios. The advantages of this lightweight network have the potential for optimization and provide strong support for solving real-world problems.

Figure 3
Comparison of different models Acc-2 and number of parameters on CMU-MOSEI



Without fine-tuning the backbone, these results show that the proposed multimodal fusion model exhibits effectiveness in sentiment analysis. It also provides f urther evidence of the need to consider semantic differences between modalities and emphasizes the importance of learning multimodal representations in the fusion process.

4.4. Ablation Study

This section conducts a series of ablation studies on the CMU-MOSEI. First, this paper conducts unimodal experiments to verify the semantic differences between the three modalities. We use text, visual, and audio data for prediction and show the experimental results in Table 4. The results show that visual and audio modalities have significant differences in F1 and Acc-2 metrics compared to textual modalities, which indicates that visual and audio modalities have lower-level semantic features than textual modalities, and there is a certain difference. Therefore, it is necessary to consider enhancing the semantic information of modalities to reduce the differences between them.

Next, this paper verifies the effect of multimodality on model performance by eliminating one modality. The

Table 4
Ablation experiments using CMU-MOSEI alignment data

Model	Corr	Acc-2	F1
Tri-CLT	**0.75**	**85.55**	**85.33**
Text	**0.71**	**83.6**	**83.3**
Visual	0.26	64.2	63.4
Audio	0.21	65.2	62.9
w/o T	0.33	70.31	69.13
w/o V	0.73	84.52	84.21
w/o A	**0.73**	**84.55**	**84.61**
A-V Fusion	0.30	69.93	70.23
T-A fusion	0.73	84.50	84.30
T-V fusion	**0.73**	**84.68**	**84.58**
w/[CLS]	0.74	85.01	85.09
w/o CL	0.74	**85.61**	**85.12**
w/o ω	0.74	85.17	84.86
w/o CMA	0.74	84.88	84.99

experimental results show that eliminating each modality leads to a decrease in the model performance, which indicates that the three modalities are essential for solving the task of MSA.

Meanwhile, it can be demonstrated through the results in Table 4 that the pairwise fusion of features from three different modalities can significantly improve the model's overall performance. Specifically, A-V Fusion, T-A Fusion, and T-V Fusion refer to Integrating Fusion Blocks to fuse different modal features for direct sentiment prediction. By comparing the performance when using audio and visual alone, it can be seen that the A-V Fusion improves by approximately 8% in the F1 value. The T-A Fusion and T-V Fusion, on the other hand, fuse audio and vision with text modality, respectively, and achieve an improvement of approximately 1.1% in F1 values when comparing the performance when text is used alone. The above analysis illustrates that incorporating the Integrating Fusion Block module during the fusion process enhances feature information, reduces semantic differences, and decreases information redundancy.

In addition, to study the effect of different parts on the performance, the following experiments were performed in this paper:

1. Removing the contrastive learning (CL) component, the results show an improvement in Acc-2 metrics but a decrease in Corr and F1 values, which proves the effectiveness of contrastive learning in multi-modal representation learning.
2. The experiment further evaluated the effect of [cls] tokens aggregation information on the model. The experiment used the output[cls] tokens from the multimodal fusion transformer as the result of sentiment prediction, and the result showed that not using [cls] tokens in the sequence processing stage was effective in predicting the result of the model.
3. w/o CMA indicates that cross-modal attention module was not used, and direct usage of fused features for sentiment analysis. The results show that using cross-modal attention to consider the complementarity and semantic distinctiveness between modalities gives better results.
4. When ω is not taken into account (concerning the case of w/o ω, where ω is set to 1), the results show that it is necessary to take ω into account in the overall loss. ω Being too large or too small affects the experimental results. In summary, better performance and information transfer capability can be obtained by these methods, which is of great significance for the further application and development of multimodal fusion technology.

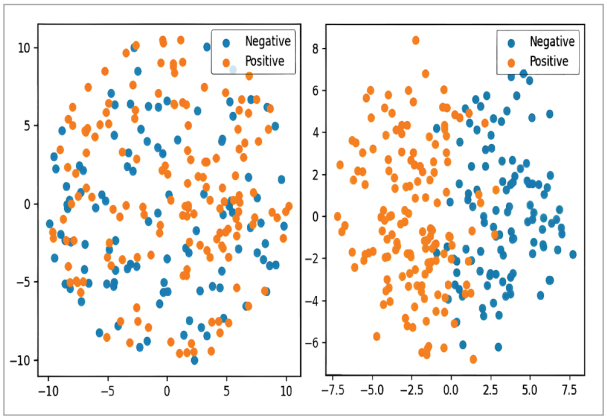
4.5. Discussion on the Selection of ω

After verifying the validity of the overall loss function parameter ω, this paper conducts further experiments to investigate the effect of different ω values on the model performance. We present the experimental results in Table 5.

Table 5
Discussion on the selection of ω on CMU-MOSEI

	F1	Corr	Acc-2	Acc-7	MAE
ω= 0.3	84.90	0.746	85.20	51.13	0.571
ω= 0.4	84.85	0.744	85.00	51.06	0.569
ω= 0.5	85.26	0.746	85.44	**52.79**	0.572
ω= 0.6	84.55	0.75	84.87	51.58	0.573
ω= 0.7	**85.33**	**0.751**	**85.55**	52.18	**0.568**
ω= 0.8	84.95	0.741	85.14	50.59	0.576

Figure 4
Visualize each batch with a set size 256 for the data in each batch



Tri-CLT performs best when ω is set to 0.7. However, the accuracy of Tri-CLT decreases when ω is set above or below 0.7. It shows that the choice of ω is vital for the model to reach the optimal solution.

4.6. Visualization

This section visualizes the multimodal embedding space of each batch of data to demonstrate the impact of contrastive learning on multimodal representation.

We show the embedding space without contrastive learning in the left panel of Figure 4 and the embedding space of the model with contrastive learning in the right panel. We selected Positive and Negative Sentiment samples from the CMU-MOSEI dataset, and multimodal representations were converted into feature points for visualization using the T-SNE algorithm. When contrastive learning is not used, the distribution of feature points tends to be very spread out, making it impossible to form distinguishable clusters. However, adding contrastive learning leads to precise categorization between feature points. It indicates that we map semantically similar inputs to similar locations, which proves beneficial for classifier prediction.

5. Conclusion

This paper proposes a transformer-based network architecture, Tri-CLT, which aims to fuse audio, visual, and textual modal features to achieve multimodal sentiment recognition. Specifically, this paper considers the semantic differences and complementarities among the three modalities. It fuses the three modalities to enhance low-level semantic features and effectively reduce the semantic differences among the modalities. Firstly, the proposed Integrating Fusion Block successfully overcomes the secondary complexity when dealing with paired sequences in the transformer, reducing computational complexity. The obtained fusion features achieve the enhancement of the three modal features. Secondly, considering the complementarity between modalities, the introduction of cross-modal attention enhances the fused features of the three modalities, achieving complementary learning between modalities. In addition, this paper introduces inter-modal contrastive learning to train the system so that the network learns a robust multimodal embedding space and improves the model's representation learning capability. Ultimately, extensive experiments on the CMU-MOSEI dataset were conducted to demonstrate the effectiveness of the proposed model.

Although Tri-CLT performs well on the CMU-MOSEI dataset, the work in this paper has some limitations. When confronted with multimodal data with different acquisition modalities, Tri-CLT may struggle to achieve the desired results, and better fusion strategies may lead to a degradation of the generalization ability on multimodal data. Future work will improve these limitations, and more modalities will be introduced for learning.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Program for Innovation Research Groups at Institutions of Higher Education in Chongqing (CXQT21032), the Fundamental Research Funds for the Natural Science Foundation of Chongqing, China (cstc2021ycjh-bgzxm0088) and the Fundamental Research Funds for the Science and Technology Research Project of Chongqing Municipal Education Commission (KJZD-M202303401).

Declaration of Interest Statement

The authors state that they have no conflicts of interest to declare.

References

1. Abdullah, S. M. A., Ameen, S. Y. A., Sadeeq, M. A., Zeebaree, S. Multimodal Emotion Recognition Using Deep Learning. Journal of Applied Science and Technology Trends, 2021, 52-58. https://doi.org/10.38094/jastt20291
2. Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., Gong, B. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021), 2021, 24206-24221.
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A. Frozen in time: A Joint Video and Image Encoder for End-to-End Retrieval. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, 1728-1738. https://doi.org/10.1109/ICCV48922.2021.00175
4. Baltrušaitis, T., Ahuja, C., Morency, L. P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 423-443. https://doi.org/10.1109/TPAMI.2018.2798607
5. Boulahia, S. Y., Amamra, A., Madi, M. R., Daikh, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications, 2021, 32-121. https://doi.org/10.1007/s00138-021-01249-8
6. Cheng, Y., Wang, R., Pan, Z., Feng, R., Zhang, Y. Look, Listen, and Attend: Co-attention Network for Self-Supervised Audio-Visual Representation Learning. MM ,20: Proceedings of the 28th ACM International Conference on Multimedia, 2020, 3884-3892. https://doi.org/10.1145/3394171.3413869
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR), Austria, 2021.
8. Fu, Z., Liu, F., Xu, Q., Qi, J., Fu, X., Zhou, A., Li, Z. NHFNET: A Non-Homogeneous Fusion Network for Multimodal Sentiment Analysis. 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 2022, 1-6. https://doi.org/10.1109/ICME52920.2022.9859836
9. Gao, T., Yao, X., Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Dominican, 2021, 6894-6910. https://doi.org/10.18653/v1/2021.emnlp-main.552
10. Gong, Y., Chung, Y. A., Glass, J. Ast: Audio Spectrogram Transformer. Interspeech, 2021, 571-575. https://doi.org/10.21437/Interspeech.2021-698
11. Han, W., Chen, H., Poria, S. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, 9180-9192. https://doi.org/10.18653/v1/2021.emnlp-main.723
12. Hazarika, D., Zimmermann, R., Poria, S. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. MM ,20: Proceedings of the 28th ACM International Conference on Multimedia, 2020, 1122-1131. https://doi.org/10.1145/3394171.3413678
13. Hori, C., Hori, T., Lee, T. Y., Zhang, Z., Harsham, B., Hershey, J. R., Sumi, K. Attention-Based Multimodal Fusion for Video Description. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, 4203-4212. https://doi.org/10.1109/ICCV.2017.450
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Krishnan, D. Supervised Contrastive Learning. Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020), 2020, 18661-18673.
15. Krishna, D. N., Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. Interspeech, 2020, 4243-4247. https://doi.org/10.21437/Interspeech.2020-1190
16. Liao, J., Zhong, Q., Zhu, Y., Cai, D. Multimodal Physiological Signal Emotion Recognition Based on Convolutional Recurrent Neural Network. IOP Conference Series: Materials Science and Engineering. 2020, 782(3). https://doi.org/10.1088/1757-899X/782/3/032005
17. Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., Morency, L. P. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. 56th Annual Meeting of the Association for Computational Linguistic, Melbourne, Australia, 2018, 2247-2256.https://doi.org/10.18653/v1/P18-1209
18. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T. CLIP4Clip: An Empirical Study of CLIP for End to

End Video Clip Retrieval and Captioning. Neurocomputing, 2022, 508, 293-304. https://doi.org/10.1016/j.neucom.2022.07.028

19. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Zhou, M. Univl: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arXiv.org, 2020, arXiv: 2002.06353.
20. Mai, S., Hu, H., Xing, S. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 481-492. https://doi.org/10.18653/v1/P19-1046
21. Mai, S., Hu, H., Xing, S. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(01), 164-172. https://doi.org/10.1609/aaai.v34i01.5347
22. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C. Attention Bottlenecks for Multimodal Fusion. Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021), 34, 14200-14213.
23. Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., Fookes, C. Deep Spatio-Temporal Feature Fusion with Compact Bilinear Pooling For Multimodal Emotion Recognition. Computer Vision and Image Understanding (CVIU), 2018, 33-42. https://doi.org/10.1016/j.cviu.2018.06.005
24. Peng, Y., Qi, J. CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning. ACM Transactions on Multimedia Computing. Communications, and Applications (TOMM), 2019, 15(1), 1-24. https://doi.org/10.1145/3284750
25. Poria, S., Cambria, E., Howard, N., Huang, G. B., Hussain, A. Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content. Neurocomputing, 2016, 174(1), 50-59. https://doi.org/10.1016/j.neucom.2015.01.095
26. Rahman, W., Hasan, M. K., Lee, S., Zadeh, A. Mao, C., Morency, L. P., Hoque, E. Integrating Multimodal Information in Large Pretrained Transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, 2359-236. https://doi.org/10.18653/v1/2020.acl-main.214
27. Shvetsova, N., Chen, B., Rouditchenko, A., Thomas, S., Kingsbury, B., Feris, R. S., Kuehne, H. Everything at Once-Multi-modal Fusion Transformer for Video Retrieval. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2022, 20020-20029. https://doi.org/10.1109/CVPR52688.2022.01939
28. Stappen, L., Baird, A., Schumann, L., Bjorn, S. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. IEEE Transactions on Affective Computing, 2021, 1334-1350. https://doi.org/10.1109/TAFFC.2021.3097002
29. Sun, Z., Sarma, P., Sethares, W., Liang, Y. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05), 8992-8999. https://doi.org/10.1609/aaai.v34i05.6431
30. Tsai, Y. H. H., Liang, P. P., Zadeh, A., Morency, L. P., Salakhutdinov, R. Learning Factorized Multimodal Representations. International Conference on Learning Representations, LA, USA, 2019.
31. Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, 6558-6569. https://doi.org/10.18653/v1/P19-1656
32. Tsai, Y. H. H., Ma, M. Q., Yang, M., Salakhutdinov, R., Morency, L. P. Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, 1823-1833. https://doi.org/10.18653/v1/2020.emnlp-main.143
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. Attention is all you need. Thirty-first Conference on Neural Information Processing Systems (NeurIPS 2017), 2017, 30, 5998-6008.
34. Yang, J., Wang, Y., Yi, R., Zhu, Y., Rehman, A., Zadeh, A., Morency, L. P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021, 1009-1021. https://doi.org/10.18653/v1/2021.naacl-main.79
35. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L. P. Tensor Fusion Network for Multimodal Sentiment Analysis. Proceedings of the 2017 Conference on Empi-

rical Methods in Natural Language Processing, Copenhagen, Denmark, Dominican, 2017, 1103-1114. https://doi.org/10.18653/v1/D17-1115

36. Zadeh, A., Zellers, R., Pincus, E., Morency, L. P. Mosi: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv.org, 2016, arXiv: 1606.06259.
37. Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., Morency, L. P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, 2236-2246. https://doi.org/10.18653/v1/P18-1208
38. Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., Morency, L. P. Memory Fusion Network for Multi-view Sequential Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1). https://doi.org/10.1609/aaai.v32i1.12021