CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network

Jiajia Tang[1]*, **Kang Li**[1]*, **Xuanyu Jin**[1], **Andrzej Cichocki**[2 3], **Qibin Zhao**[4], **Wanzeng Kong**[1]†
[1]Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province,
School of Computer Science and Technology, Hangzhou Dianzi University, China
[2]Skolkovo Institute of Science and Technology, Moscow, Russia
[3]Systems Research Institute, Polish Academy of Science, Warsaw, Poland
[4]Center for Advanced Intelligence Project, RIKEN
{hdutangjiajia, Jxuanyu599}@163.com
{likang_bro, kongwanzeng}@hdu.edu.cn
{a.cichocki, qibin.zhao}@riken.jp

Abstract

Multimodal sentiment analysis is the challenging research area that attends to the fusion of multiple heterogeneous modalities. The main challenge is the occurrence of some missing modalities during the multimodal fusion procedure. However, the existing techniques require all modalities as input, thus are sensitive to missing modalities at predicting time. In this work, the coupled-translation fusion network (CTFN) is firstly proposed to model bi-direction interplay via couple learning, ensuring the robustness in respect to missing modalities. Specifically, the cyclic consistency constraint is presented to improve the translation performance, allowing us directly to discard decoder and only embraces encoder of Transformer. This could contribute to a much lighter model. Due to the couple learning, CTFN is able to conduct bi-direction cross-modality intercorrelation parallelly. Based on CTFN, a hierarchical architecture is further established to exploit multiple bi-direction translations, leading to double multimodal fusing embeddings compared with traditional translation methods. Moreover, the convolution block is utilized to further highlight explicit interactions among those translations. For evaluation, CTFN was verified on two multimodal benchmarks with extensive ablation studies. The experiments demonstrate that the proposed framework achieves state-of-the-art or often competitive performance. Additionally, CTFN still maintains robustness when considering missing modality.

1 Introduction

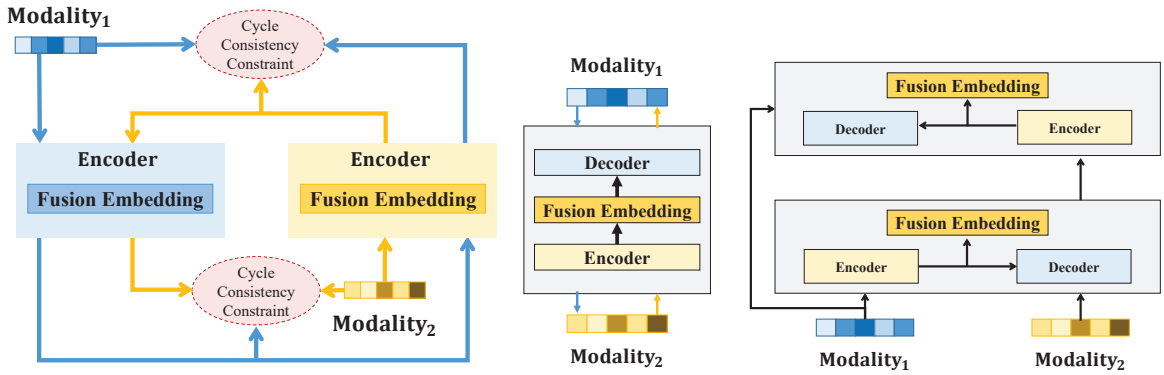
Sentiment analysis has witnessed many significant advances in the artificial intelligence community, in which text (Yadollahi et al., 2017), visual (Kahou et al., 2016), and acoustic (Luo et al., 2019) modalities are primarily employed to the related research respectively, allowing to exploit the human emotional characteristic and intention effectively (Deng et al., 2018). Intuitively, due to the consistency and complementarity among different sources, the joint representation attend to reason about multimodal messages, which are capable of boosting the performance of the specific task (Pan et al., 2016; Gebru et al., 2017; Al Hanai et al., 2018).

Multimodal fusion procedure is to incorporate multiple knowledge for predicting a precise and proper outcome (Baltrušaitis et al., 2018). Historically, the existing fusion has been done generally by leveraging the model-agnostic process, considering the early fusion, late fusion, and hybrid fusion technique (Poria et al., 2017a). Among those, early fusion focussed on the concatenation of the unimodal presentation (D'mello and Kory, 2015). On the contrast, late fusion performs the integration at the decision level, by voting among all the model results (Shutova et al., 2016). As to the hybrid fusion, the output comes from the combination of the early fusion and unimodal prediction (Lan et al., 2014). Nevertheless, multimodal sentiment sequences often consists of unaligned properties, and the traditional fusion manners are failed to take the heterogeneity and misalignment into account carefully, which raises a question on investigating the more sophisticated models and estimating emotional information. (Tsai et al., 2020; Niu et al., 2017).

Recently, Transformer-based multimodal fusion framework has been developed to address the above issues with the help of multi-head attention mechanism (Rahman et al., 2020; Le et al., 2019; Tsai et al., 2019). By introducing the standard Transformer network (Vaswani et al., 2017) as the basis, Tsai *et al.* (Tsai et al., 2019) captured the integrations directly from unaligned multimodal streams in an end-to-end fashion, latently adapted streams from one modality to another with the cross-modal

*Equal contribution
†Corresponding author: Wanzeng Kong

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5301–5311
August 1–6, 2021.



(a) CTFN model. (b) MCTN model. (c) TransModality model.

Figure 1: Comparison of CTFN with existing translation-based models. In our model, the cyclic consistency constraint is presented to improve the translation performance, allowing us directly to discard decoder and only embrace encoder of Transformer. This could contribute to a much lighter model. Due to the couple learning, CTFN is able to conduct bi-direction cross-modality intercorrelation parallelly, ensuring the robustness in respect to missing modalities.

attention module, regardless of the need for alignment. Furthermore, Wang *et al.* (Wang et al., 2020) proposed a parallel Transformer unit, allowing to explore the correlation between multimodal knowledge effectively. However, the decoder component of standard Transformer is employed to improve the translation performance, which may lead to some redundancy. Moreover, the explicit interaction among cross-modality translations were not considered. Essentially, compared to our CTFN, their architecture require access to all modalities as inputs for exploring multimodal interplay with the sequential fusion strategy, thus are rather sensitive in the case of multiple missing modalities.

In this paper, CTFN is proposed to model bi-directional interplay based on coupled learning, ensuring the robustness in respect to missing modalities. Specifically, the cyclic consistency constraint is proposed to improve the translation performance, allowing us directly to discard decoder and only embrace encoder of Transformer. This could contribute to a much lighter model. Thanks to the couple learning, CTFN is able to conduct bi-direction cross-modality intercorrelation parallelly. Take CTFN as a basis, a hierarchical architecture is established to exploit modality-guidance translation. Then, the convolution fusion block is presented to further explore the explicit correlation among the above translations. Importantly, based on the parallel fusion strategy, our CTFN model still provides flexibility and robustness when considering only one input modality.

For evaluation, CTFN was verified on two multimodal sentiment benchmarks, CMU-MOSI (Zadeh et al., 2016) and MELD (Poria et al., 2019). The experiments demonstrate that CTFN could achieve the state-of-the-art or even better performance compared to the baseline models. We also provide several extended ablation studies, to investigate intrinsic properties of the proposed model.

2 Related Work

The off-the-shelf multimodal sentiment fusion architecture comprises two leading groups: translation-based and non-translation based model.

Non-translation based: Recently, RNN-based models, considering GRU and LSTM, have received significant advances in exploiting the context-aware information across the data (Yang et al., 2016; Agarwal et al., 2019). $bc-LSTM$ (Poria et al., 2017b) and $GME-LSTM$ (Chung et al., 2014) presented a LSTM-based model to retrieve contextual information, where the unimodal features are concatenated into a unit one as the input information. Similarly, $MELD-base$ (Poria et al., 2019) leveraged the concatenation of audio and textual features on the input layer, and employed GRU to model sentimental context. In contrast, $CHFusion$ (Majumder et al., 2018) employed the RNN-based hierarchical structure to draw fine-grained local correlations among the modalities, and the empirical evidence illustrates superior advances compared to the simple concatenation of unimodal presentation. On the basis of RNN, $MMMU-BA$ (Ghosal et al., 2018) further employed multimodal attention block to absorb the contribution of all the neighboring utterances, which demonstrates that the attention mechanism

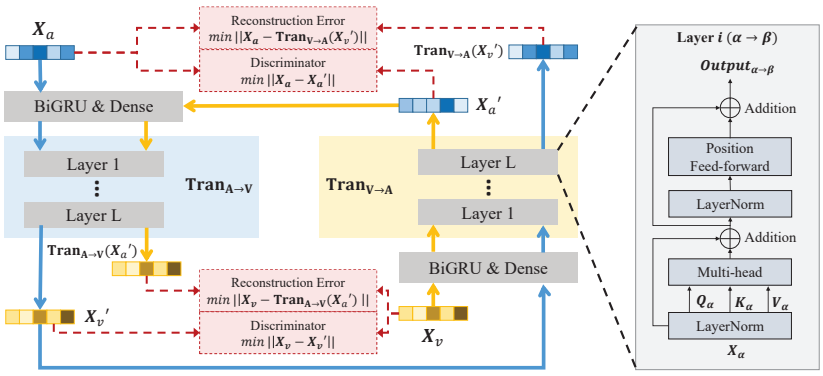


Figure 2: CTFN: $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$ refer to the features of modality *audio* and *video* respectively. The blue line indicates the primal process, and the yellow line indicates the dual procedure. Note that the cyclic consistency constraint is presented to improve the translation performance, allowing us directly to discard decoder and only embrace encoder of Transformer. And thanks to couple learning, CTFN could combine primal and dual process into a coupled structure, ensuring the robustness in respect to missing modalities.

can utilize the neighborhood contribution for integrating the contextual information. However, all these methods are suitable for the low-level presentation within the single modality with a non-translation manner, which may be easily sensitive to the noisy terms and missing information in the sources.

Translation-based model: Inspired by the recent success of sequence to sequence (Seq2Seq) models (Lin et al., 2019; **?**) in machine translation, (Pham et al., 2019) and (Pham et al., 2018) presented multimodal fusion model via the essential insight that translates from a source modality to a target modality, which is able to capture much more robust associations across multiple modalities. $MCTN$ model incorporated a cyclic translation module to retrieve the robust joint representation between modalities in a sequential manner, e.g., the language information firstly associated with the visual modality, and latently translated into the acoustic modality. Compared with the $MCTN$, $Seq2Seq2Sent$ introduced a hierarchical fusion model using the Seq2Seq methods. For the first layer, the joint representation of a modality pair is treated as an input sequence for the next Seq2Seq layer in an attempt to decode the third modality. Inspired by the success of the Transformer-based model, Tsai *et al.* introduced a directional cross-modality attention module to extend the standard Transformer network. Follow the basic idea of Tsai *et al.*, Wang *et al.* provided a novel multimodal fusion cell which is comprised of two standard Transformers, embracing the association with a modality pair during the forward and backward translation implicitly. However, all existing models adopt sequential multimodal fusion architecture, which requires all modalities as input, therefore they can be sensitive to the case of multiple missing modalities. Moreover, the explicit interactions among cross-modality translations were not considered.

3 Methodology

In this section, we firstly present CTFN (Figure 2), which is capable of exploring bi-direction cross-modality translation via couple learning. On the basis of CTFN, a hierarchical architecture is established to exploit multiple bi-direction translations, leading to double multimodal fusing embeddings (Figure 4). Then, the convolutional fusion block (Figure 3) is applied to further highlight explicit correlation among cross-modality translations.

3.1 Preliminaries

The two benchmarks consist of three modalities, audio, video and textual modality. Specifically, the above utterance-level modalities are denoted as $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times d_a}$, $\boldsymbol{X_v} \in \mathbb{R}^{T_v \times d_v}$ and $\boldsymbol{X_t} \in \mathbb{R}^{T_t \times d_t}$, respectively. The number of utterances is presented as $T_i (i \in \{a, v, t\})$, and $d_i (i \in \{a, v, t\})$ stands for the dimension of the unimodality features.

3.2 Coupled-Translation Fusion Network

For simplicity, we consider two unimodality presentation $\boldsymbol{X_a}$ and $\boldsymbol{X_v}$ explored from audio (A) and video (V), respectively. In the primal process of CTFN, we focus on learning a directional translator $Tran_{A\to V}(\boldsymbol{X_a}, \boldsymbol{X_v})$ for translating the modality audio to video. Then, the dual process aims to learn an inverse directional translator $Tran_{V\to A}(\boldsymbol{X_v}, \boldsymbol{X_a})$, allowing for the translation from modality video to audio. Inspired by the suc-

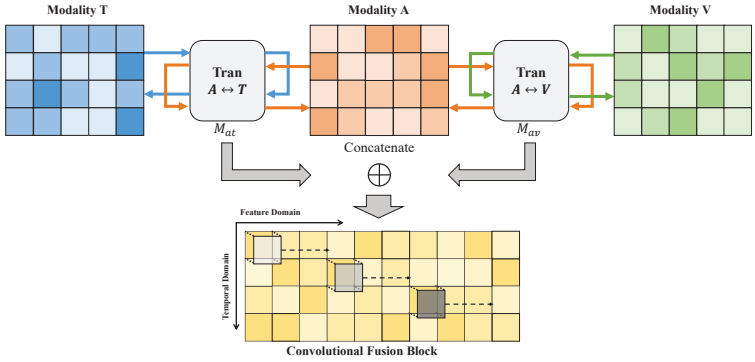


Figure 3: Multimodal convolutional fusion block: $\boldsymbol{M_{at}} \in \mathbb{R}^{T \times F_{at}}$ and $\boldsymbol{M_{av}} \in \mathbb{R}^{T \times F_{av}}$ refer to the cross-modality translations, where T and F_* are the size of time and feature domain respectively. Subsequently, $\boldsymbol{M_{at}}$ and $\boldsymbol{M_{av}}$ are concatenated along the feature domain, and the convolution operation is utilized to exploit the local and explicit interplay between cross-modality translations.

cess of Transformer in Natural Language Processing, the encoder of Transformer is introduced to our model as the translation block, which is an efficient and adaptive manner for retrieving the long-range interplay along the temporal domain. Importantly, the cyclic consistency constraint is presented to improve the translation performance. And due to the couple learning, CTFN is able to combine primal and dual process into a coupled structure, ensuring the robustness in respect to missing modalities.

For the primal task, $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times d_a}$ is firstly delivered to a densely connected layer for receiving a linear transformation $\boldsymbol{X_a} \in \mathbb{R}^{T_a \times L_a}$, where L_a is the output dimension of the linear layer. And the corresponding query matrix, key matrix and value matrix are denoted as $\boldsymbol{Q_a} = \boldsymbol{X_a}\boldsymbol{W_{Q_a}} \in \mathbb{R}^{T_a \times L_a}$, $\boldsymbol{K_a} = \boldsymbol{X_a}\boldsymbol{W_{K_a}} \in \mathbb{R}^{T_a \times L_a}$, $\boldsymbol{V_a} = \boldsymbol{X_a}\boldsymbol{W_{V_a}} \in \mathbb{R}^{T_a \times L_a}$, where $\boldsymbol{W_{Q_a}} \in \mathbb{R}^{L_a \times L_a}$, $\boldsymbol{W_{K_a}} \in \mathbb{R}^{L_a \times L_a}$ and $\boldsymbol{W_{V_a}} \in \mathbb{R}^{L_a \times L_a}$ are weight matrixes. The translation from modality A to V is performed as $\boldsymbol{X_v}' = Tran_{A \to V}(\boldsymbol{X_a}, \boldsymbol{X_v}) \in \mathbb{R}^{T_a \times L_v}$, where $\boldsymbol{X_v}'$ refers to the fake $\boldsymbol{X_v}$, and $\sqrt{L_a}$ is the scale coefficient. Note that the input $\boldsymbol{X_a}$ is directly delivered to the translation process, while the input $\boldsymbol{X_v}$ is used to analyze the difference between real data $\boldsymbol{X_v}$ and fake output $\boldsymbol{X_v}'$. Subsequently, $\boldsymbol{X_v}'$ is passed through the $Tran_{V \to A}$, leading to the reconstruct output $\boldsymbol{X_a}' = Tran_{V \to A}(\boldsymbol{X_v}', X_a)$, and the $\boldsymbol{X_a}$ is only used to calculate the diversity between the real and reconstruct data.

$$\begin{aligned}\boldsymbol{X_v}' &= Tran_{A \to V}(\boldsymbol{X_a}, \boldsymbol{X_v}) \\ &= softmax(\frac{\boldsymbol{Q_a}{\boldsymbol{K_a}}^T}{\sqrt{L_a}})\boldsymbol{V_a} \\ &= softmax(\frac{\boldsymbol{X_a}\boldsymbol{W_{Q_a}}{\boldsymbol{W_{K_a}}}^T{\boldsymbol{X_a}}^T}{\sqrt{L_a}})\boldsymbol{X_a}\boldsymbol{W_{V_a}}. \quad (1)\end{aligned}$$

Analogously, in the dual process, $\boldsymbol{X_v} \in \mathbb{R}^{T_v \times L_v}$ is captured based on the input $\boldsymbol{X_v} \in \mathbb{R}^{T_v \times d_v}$, $\boldsymbol{X_a}' = Tran_{V \to A}(\boldsymbol{X_v}, \boldsymbol{X_a}) \in \mathbb{R}^{T_a \times L_a}$, and reconstructed representation $\boldsymbol{X_v}' = Tran_{A \to V}(\boldsymbol{X_a}', \boldsymbol{X_v}) \in \mathbb{R}^{T_v \times L_v}$. Essentially, $Tran_{A \to V}$ and $Tran_{V \to A}$ are implemented by several sequential encoder layers. During the translation period, we hypothesize that intermediate encoder layer contains the cross-modality fusion information and effectively balance the contribution of two modalities. Hence, the output of the middle encoder layer ${Tran_{A \to V}}^{[L/2]}$ and ${Tran_{V \to A}}^{[L/2]}$ stand for the multimodal fusion knowledge, where L refers to the number of layers, and when L is odd number, then $L = L + 1$.

As for the model reward, the primal process has an immediate reward $r_p = \|\boldsymbol{X_a} - Tran_{V \to A}(\boldsymbol{X_v}')\|_F$, and the dual step related reward is $r_d = \|\boldsymbol{X_v} - Tran_{A \to V}(\boldsymbol{X_a}')\|_F$, indicating the similarity between the real data and the reconstructed output of the translator. For simplicity, a linear transformation module is adopted to combine the primal and dual step reward into a total model reward, e.g., $r_{all} = \alpha r_p + (1 - \alpha) r_d$, where α is employed to balance the contribution between dual and primal block. Additionally, the loss functions utilized in the coupled-translation multimodal fusion block are defined as follows:

$$\begin{aligned}l_{A \to V}(\boldsymbol{X_a}, \boldsymbol{X_v}) =& \|Tran_{A \to V}(\boldsymbol{X_a}, \boldsymbol{X_v}) - \boldsymbol{X_v}\|_F + \\ & \|Tran_{A \to V}(\boldsymbol{X_a}', \boldsymbol{X_v}) - \boldsymbol{X_v}\|_F \\ l_{V \to A}(\boldsymbol{X_v}, \boldsymbol{X_a}) =& \|Tran_{V \to A}(\boldsymbol{X_v}, \boldsymbol{X_a}) - \boldsymbol{X_a}\|_F + \\ & \|Tran_{V \to A}(\boldsymbol{X_v}', \boldsymbol{X_a}) - \boldsymbol{X_a}\|_F \\ l_{A \leftrightarrow V} =& \alpha l_{A \to V}(\boldsymbol{X_a}, \boldsymbol{X_v}) + (1 - \alpha) l_{V \to A}(\boldsymbol{X_v}, \boldsymbol{X_a}),\end{aligned} \quad (2)$$

where $l_{A \to V}(\boldsymbol{X_a}, \boldsymbol{X_v})$ and $l_{V \to A}(\boldsymbol{X_v}, \boldsymbol{X_a})$ refer to the training loss of the primal and dual translator respectively, and $l_{A \leftrightarrow V}$ stands for the loss of bi-directional translator unit. Essentially, when the training process of all coupled-translation blocks

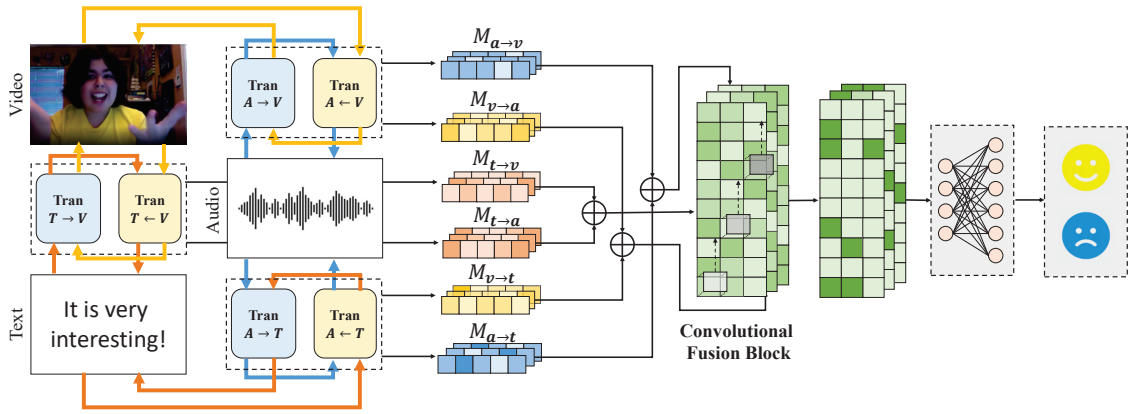


Figure 4: The hierarchical framework associated with three CTFNs during the training period. Each CTFN is utilized to explore the specific bi-direction cross-modality interplay. On the basis of this, three CTFN are stacked into a united one for exploiting multiple bi-direction translations, leading to double multimodal fusing embeddings. Then, multiple multimodal fusing embeddings are delivered to the multimodal convolutional fusion block.

are finished, our model only needs one input modality at predicting time, without the help of target modalities.

Indeed, $l_{A\leftrightarrow V}$ indicates the cycle-consistency constraint in our couple learning model. The cycle-consistency is well-known, which refers to combination of forward and backward cycle-consistency. However, our goal is to solve missing modality problem in multi-modal learning, which cannot be achieved by applying cycle-consistency straightforward. This is because that introducing this strict cycle-consistency to CTFN fail to effectively associate primal task with dual task of the couple learning model. To solve this problem, we relaxed constraint of original cycle-consistency by using a parameter 'α' to balance the contribution of forward and backward cycle-consistency, leading to a much more flexible cycle-consistency. Thanks to the great flexibility of new proposed cycle-consistency, we could adaptively and adequately associate primal with dual task, resulting in much more balanced consistency among modalities.

3.3 Multimodal convolutional fusion block

Based on CTFN, each modality is treated as the source moment for $(M-1)$ times, which means that each modality holds $(M-1)$ directional translations, $\{Tran_{modality\_source\to modality\_m}\}_{m=1}^{M}$, where M refers to the total number of modalities. For instance, given modality audio, we can retrieve the following two modality-guidance translations:

$$[Tran_{a\to v}{}^{L/2}, video'] = Tran_{a\to v}(audio, video)$$
$$[Tran_{a\to t}{}^{L/2}, text'] = Tran_{a\to t}(audio, text). \quad (3)$$

Note that audio plays a key role in different cross-modality translations, and provides the strong guidance for capturing various cross-modality interplay. For blending the contribution of source modality (audio) effectively, a convolution fusion block is incorporated to explore explicit and local correlation among modality-guidance translations.

Initially, the two cross-modality intermediate correlations $Tran_{audio\to vedio}{}^{L/2}$ and $Tran_{audio\to text}{}^{L/2}$ are concatenated along the temporal domain into a unit representation, where the size of time sequence is equal ($T_a = T_v = T_t$), thus the concatenation is of size $T_a \times (L_v + L_t)$:

$$Z_{concat} = Tran_{a\to v}{}^{L/2} \oplus Tran_{a\to t}{}^{L/2}. \quad (4)$$

Subsequently, the temporal convolution is employed to further retrieve explicit interactions among cross-modality translations. Specifically, we adopt a 1D temporal convolutional layer to exploit the local patten in a light manner:

$$\hat{Z}_{concat} = Conv1D(Z_{concat}, K_{concat}) \in \mathbb{R}^{T_a \times L_d}, \quad (5)$$

where K_{concat} is the size of the convolutional kernel, and L_d is the length of the cross-modality integration dimension. The temporal kernel is used to perform the convolutional operation along the feature dimension, allowing to further exploit local interplay among cross-modality translations. That is to say, the local interplay fully exploits the contribution from modality-guidance translations.

3.4 Hierarchical Architecture

On the basis of CTFN and convolutional multimodal fusion network, a hierarchical architecture was proposed for exploiting multiple bi-direction translations, leading to double multimodal fusing embeddings. For instance, given M modalities, our model could achieve double

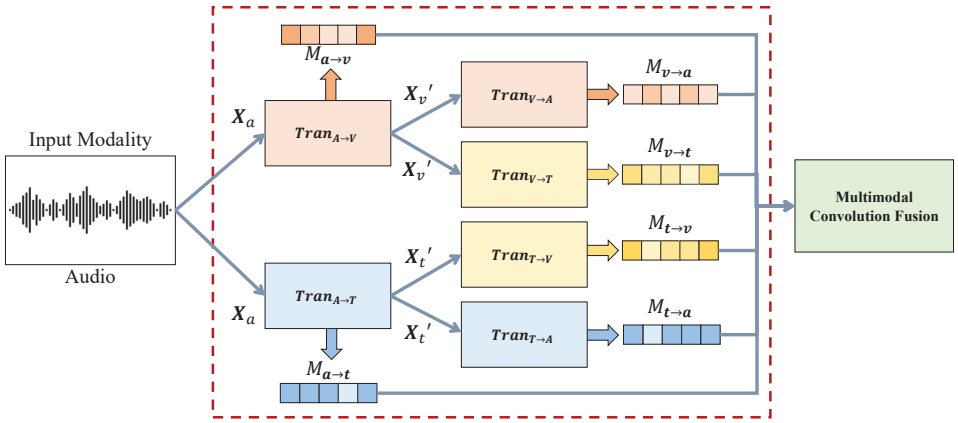


Figure 5: We only employ a single input modality (audio) to do the multimodal fusion task during the predicting period. Initially, audio presentation $\boldsymbol{X_a}$ is sent to the pre-trained translators $Tran_{A\to V}$ and $Tran_{A\to T}$ for retrieving $\boldsymbol{X_v}$' and $\boldsymbol{X_t}$'. Then, $\boldsymbol{X_v}$' is transmitted to $Tran_{V\to A}$ and $Tran_{V\to T}$ respectively. And, $\boldsymbol{X_t}$' is sent to $Tran_{T\to V}$ and $Tran_{T\to A}$ respectively. Hence, the tree structure only need one input modality to do the multimodal fusion task.

C_M^2 embeddings. As illustrated in Figure 4, the proposed architecture consists of three CTFNs $Tran_{A\leftrightarrow V}$, $Tran_{A\leftrightarrow T}$ and $Tran_{V\leftrightarrow T}$. Considering the contribution of the guidance (source) modality, the modality-guidance translations are denoted as $Tran_{T\leftarrow A\to V} = [Tran^{L/2}_{A\to V}, Tran^{L/2}_{A\to T}]$, $Tran_{T\leftarrow V\to A} = [Tran^{L/2}_{V\to T}, Tran^{L/2}_{V\to A}]$, and $Tran_{A\leftarrow T\to V} = [Tran^{L/2}_{T\to A}, Tran^{L/2}_{T\to V}]$, respectively. Similarly, when taking the contribution of target modalities into account, corresponding modality-guidance translations are illustrated as $Tran_{T\to A\leftarrow V} = [Tran^{L/2}_{V\to A}, Tran^{L/2}_{T\to A}]$, $Tran_{T\to V\leftarrow A} = [Tran^{L/2}_{T\to V}, Tran^{L/2}_{A\to V}]$, and $Tran_{A\to T\leftarrow V} = [Tran^{L/2}_{A\to T}, Tran^{L/2}_{V\to T}]$, respectively. Subsequently, the convolutional fusion layer is used to further exploit explicit local interplay among modality-guidance translations associated with the same source/target modality, which can fully leverage the contribution of source/target modality.

Essentially, as demonstrated in Figure 4, our model has "12+1" loss constraints in total, which includes 3 CTFNs, each one has 4 training loss (primal & dual translator training loss), and 1 classification loss. However, we do not need to balance these targets together, which is achieved by our training strategy that 3 CTFNs are trained individually. For each CTFN, one hyper-parameter 'α' is introduced to balance the loss of primal translator and dual translator, and this hyper-parameter is shared among 3 CTFNs. Hence, 3 CTFNs only need 1 hyper-parameter to balance the training loss, which is easy to be tuned. The classification loss is used for training the classifier on the 3 CTFNs's outputs.

4 Experiments

4.1 Experimental setups

Datasets. CMU-MOSI consists of 2199 opinion video clips from online sharing websites (e.g., YouTube). Each utterance of the video clip is annotated with a specific sentimental label of positive or negative in the range scale of $[-3, +3]$. The corresponding training, validation, and testing size refer to division set $(1284, 229, 686)$. Additionally, the same speaker will not appear in both training and testing sets, allowing to exploit speaker-independent joint representations. MELD dataset contains 13000 utterances from the famous TV-series $Friends$. Each utterance is annotated with emotion and sentiment labels, considering 7 classes of emotion tag (anger, disgust, fear, joy, neutral, sadness, and surprise) and 3 sentimental tendency levels (positive, neutral, and negative). Hence, the original dataset can be denoted as MELD (Sentiment) and MELD (Emotion) with respect to the data annotation, we only verified our model on the MELD (Sentiment). Note that CMU-MOSI and MELD are the public and widely-used datasets which have been aligned and segmented already.

Features. For CMU-MOSI dataset, we adopt the same preprocess manner mentioned in MFN (Zadeh et al., 2018) to extract the low-level representation of multimodal data, and synchronized at the utterance level that in consistent with text modality. For MELD benchmark, we follow

Models	CMU-MOSI				MELD (Sentiment)
	Bi-modality			Tri-modality	Bi-modality
	(video, audio)	(text, video)	(text, audio)	(text, audio, video)	(text, audio)
GME-LSTM (Chung et al., 2014)	52.90	74.30	73.50	76.50	66.46
bc-LSTM (Poria et al., 2017b)	56.52	78.59	78.86	79.26	66.09
MELD-based (Poria et al., 2019)	54.79	76.60	76.99	79.19	66.68
CHFusion (Majumder et al., 2018)	54.49	74.77	78.54	76.51	65.85
MMMU-BA (Ghosal et al., 2018)	57.45	80.85	79.92	81.25	65.56
SeqSeq2Sent (Pham et al., 2018)	58.00	67.00	66.00	70.00	63.84
MCTN (Pham et al., 2019)	53.10	76.80	76.40	79.30	66.27
TransModality (Wang et al., 2020)	59.97	80.58	81.25	82.71	67.04
CTFN (ours, L=1)	62.20	80.49	81.4	80.18	**67.82**
CTFN (ours, L=3)	63.11	**81.55**	**82.16**	**82.77**	67.78
CTFN (ours, L=6)	**64.48**	80.79	81.71	81.10	67.24

Table 1: Comparison of performance results for sentiment analysis on CMU-MOSI and MELD (Sentiment) benchmark using various SOTA models.

the related work of MELD, in which the 300-dimensional GloVe (Pennington et al., 2014) text vectors are fed into a 1D-CNN (Chen et al., 2017) layer to extract textual representation, and audio-based descriptors are explored with the popular toolkit openSMILE (Eyben et al., 2010), while visual features were not taken into account for the sentiment analysis.

Comparisons. We introduced the translation-based and non-translation based models to this work as the baselines. Translation-based: Multimodal Cyclic Translation Network (MCTN), Sequence to Sequence for Sentiment (Seq2Seq2Sent), Multimodal Sentiment Analysis with Transformer (TransModality). And non-translation based: bi-directional contextual LSTM (bc-LSTM), Gated Embedding LSTM (GME-LSTM), Multimodal EmotionLines Dataset baseline model (MELD-base), Hierarchical Fusion with Context Modeling (CHFusion), Multi-Modal Multi-Utterance - Bi-Modal Attention (MMMU-BA).

4.2 Experiment results and analysis

Performance comparison with state-of-the-art models. Firstly, we analyzed the performance between state-of-the-art baselines and our proposed model. The bottom rows in Table 1 indicate the effectiveness and superiority of our model. Particularly, on CMU-MOSI dataset, CTFN exceeded the previous best TransModality on (video, audio) by a margin of 4.51. Additionally, on MELD (Sentiment) dataset, the empirical improvement of CTFN was 0.78. It is interesting to note that the improvement of (video, audio) is more significant than (text, video) and (text, audio). This implies that coupled-translation structure is capable of decreasing the risk of interference between video and audio efficiently, and further leverage the explicit consistency between auxiliary features. As for (text, audio, video), CTFN exceeds the previous best TransModality with an improvement of 0.06, leading to a comparable performance. Indeed, for the same tri-modality fusion task, TransModality needs 4 encoders and 4 decoders, while CTFN only requires 6 encoders. It should be emphasized that the cyclic consistency mechanism could contribute to a much lighter model, as well as the more effective bi-directional translation. In addition, compared to the bi-modality setting, the tri-modality case achieved the improvement of 0.61, indicating the benefits brought by hierarchical architecture and convolution fusion.

Settings		CMU-MOSI			
		CTFN		SeqSeq2Sent	
		F1	Acc	F1	Acc
1 missing modality	(~~audio~~, video, text)	81.82	81.55	67.00	67.00
	(audio, ~~video~~, text)	82.23	82.16	65.00	66.00
	(audio, video, ~~text~~)	66.79	61.59	58.00	58.00
2 missing modalities	(~~audio~~,~~video~~, text)	80.78	80.79	76.00	77.00
	(audio, ~~video~~, ~~text~~)	62.82	61.43	56.00	56.00
	(~~audio~~, video, ~~text~~)	63.94	60.98	48.00	57.00
0 missing modality	(text, audio, video)	82.85	82.77	66.00	70.00

Table 2: Multimodal fusion results of SeqSeq2Sent and CTFN with missing modalities. The setting (audio, ~~video~~, ~~text~~) refers to the process that CTFN only employs a single input modality (audio) to do the multimodal fusion task, shown in Figure 5.

Effect of CTFN with missing modalities. Existing translation-based manners focus only on the join representation between modalities, and ignore the potential occurrence of the missing modalities. Therefore, we analyzed how does missing modality may affect the final performance of CTFN and the sequential translation-based model SeqSeq2Sent. Note that SeqSeq2Sent only employs LSTM to analyze uni-modality rather than the translation-based method. Specifically, we take the hierarchical architecture combined with three CTFNs as

the testing model. From the Table 2, we observe that compared to the setting (text, audio, video), the text-based settings {(~~audio~~, video, text), (audio, ~~video~~, text), (~~audio~~,~~video~~, text)} seem to reach the comparable result with only a relatively small performance drop. On the contrast, when text was missing, the model has a relatively large performance drop, which implies that language modality contains much more discriminative sentimental message than audio and video, leading to the significantly better performance. Essentially, the performance of (~~audio~~,~~video~~, text) demonstrates that hierarchical CTFN is able to maintain robustness and consistency when considering only a single input modality. In other words, the cyclic consistency mechanism allows CTFN to fully exploit the cross-modality interplay, thus hierarchical CTFN could transmit the single modality to various pre-trained CTFNs for retrieving multimodal fusion message.

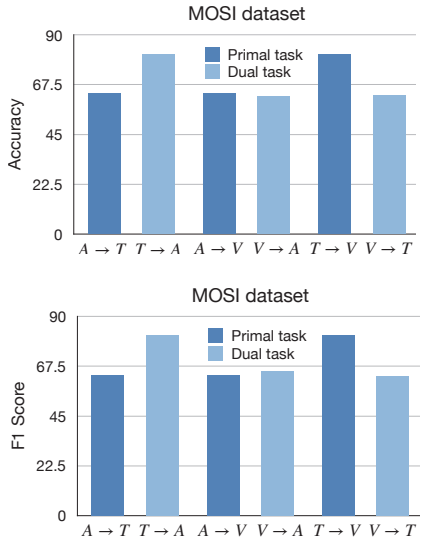


Figure 6: Effect of the translation direction.

Effect of the translation direction. In this paper, we propose a coupled-translation block, which aims to embrace fusion messages from the bi-directional translation process. Hence, we are interested to investigate the impact of translation direction. Figure 6 depicts the performance of various translations, considering (audio, text), (audio, video), and (text, video) translation. For the (audio, text) instance, the translation text→audio achieves better performance than audio→text . Similarly, the translation text→video surpasses the result of video→text. However, the performance of audio→video and video→audio seems to be quite similar. The superiority of text→video and text→audio may demonstrate that text modality possesses much more sentimental information. Moreover, the prospects of text modality allow text to be the strong backbone of the translation.

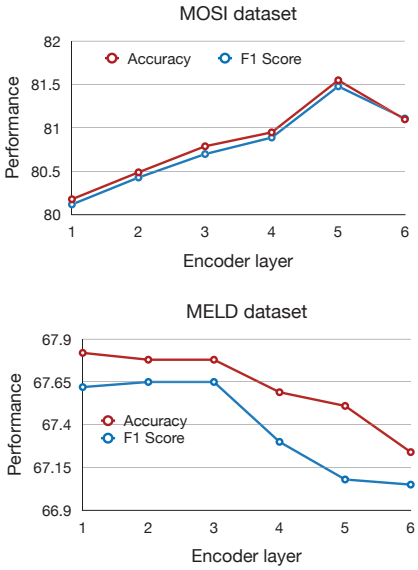


Figure 7: Effect of the translator layer.

Effect of the translator layer. As each translator is comprised of several sequential encoder layers. In this part, we assume that the output representation of a specific layer may affect the performance of the proposed model. For simplicity, we perform the related task on CMU-MOSI with the setting of (a, v, t), as well as the (t, a) on MELD (Sentiment). Initially, we retrieve the embedding from the specific layer, where the layer ranges from 1 to L (L is the total number of the layer). In Figure 7, it is interesting to note that the model reaches the peak value at $layer\_5$ on CMU-MOSI, which means that the output of the fifth layer embraces the most discriminative fusion message. In comparison, on MELD (Sentiment), the model achieves the best performance at $layer\_1$, which may imply that the simple translator associated with only one layer is able to capture the joint representation for the simple case (text, audio). In conclusion, the lower encoder layer may involve low-level characteristics of interplay, while the higher encoder layer may embrace the explicit messages. Additionally, the output of the specific layer of the encoder lies on the corresponding task and dataset. We tried also (text, audio) on MOSI, and CTFN maximizes the performance at $layer\_3$. Compared to (text, audio, video), (text, audio) is the relatively simple case, thus the lower encoder layer may is sufficient to demonstrate the interaction between text and audio.

Effect of concatenation strategy of translation. In our work, those translations associated

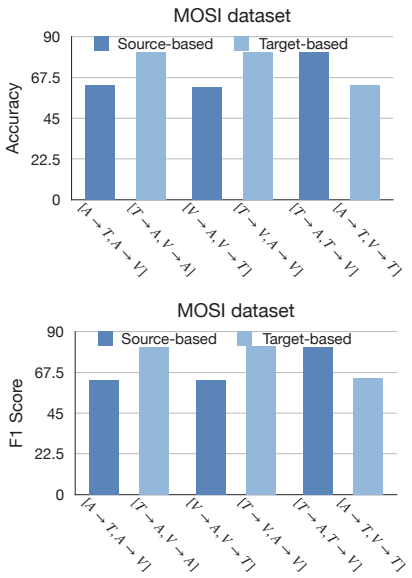


Figure 8: Effect of concatenation strategy via source/target modality on MOSI. [A→T,A→V] indicates the audio-based source concatenation [(A→T)⊕(A→V)], and [T→A,V→A] indicates the audio-based target concatenation.

with the same guidance (source) modality are concatenated along the feature domain. As each modality serves as the source and target modality in turn, we are interested to analyze the impact of the distinct concatenation strategies, e.g., concatenate the translations via the same source or target modality. As shown in Figure 8, it is obvious to find that audio-based target concatenation [(T→A) ⊕ (V→A)] performs significantly better than [(A→T)⊕(A→V)] with a large margin. Analogously, video-based target concatenation [(T→V)⊕(A→V)] works better than [(V→A)⊕(V→T)]. The above performance may indicate that joint presentation is able to achieve the significantly improved benefits with the help of guidance modality text. In conclusion, when text modality serves as the guidance modality, which may effectively leverage the contribution from audio and video, and further boost the task performance in a robust and consistent way.

5 Conclusion

In this paper, we present a novel hierarchical multimodal fusion architecture using coupled-translation fusion network (CTFN). Initially, CTFN is utilized for exploiting bi-directional interplay via couple learning, ensuring the robustness in respect to missing modalities. Specifically, the cyclic mechanism directly discards the decoder and only embraces the encoder of Transformer, which could contribute to a much lighter model. Due to the couple learning, CTFN is able to conduct bi-direction cross-modality intercorrelation parallelly. Based on CTFN, a hierarchical architecture is further established to exploit multiple bi-direction translations, leading to double multimodal fusing embeddings compared with traditional translation methods. Additionally, a multimodal convolutional fusion block is employed to further explore the complementarity and consistency between cross-modality translations. Essentially, the parallel fusion strategy allows the model maintains robustness and flexibility when considering only one input modality. CTFN was verified on two public multimodal sentiment benchmarks, the experiments demonstrate the effectiveness and flexibility of CTFN, and CTFN achieves state-of-the-art or comparable performance on CMU-MOSI and MELD (Sentiment). For future work, we like to evaluate CTFN on more multimodal fusion tasks. The source code can be obtained from https://github.com/deepsuperviser/CTFN.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and valuable suggestions. This work was supported by National Key R&D Program of China for Intergovernmental International Science and Technology Innovation Cooperation Project (2017YFE0116800), National Natural Science Foundation of China (U20B2074, U1909202), Science and Technology Program of Zhejiang Province (2018C04012), Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province (2020E10010), JSPS KAKENHI (Grant No. 20H04249), and supported by the Ministry of Education and Science of the Russian Federation (Grant 14.756.31.0001).

References

Ayush Agarwal, Ashima Yadav, and Dinesh Kumar Vishwakarma. 2019. Multimodal sentiment analysis via rnn variants. In *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, pages 19–23.

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learn-

ing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Tao Chen, Ruifeng Xu, Yulan He, and Xuan Wang. 2017. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Didan Deng, Yuqian Zhou, Jimin Pi, and Bertram E Shi. 2018. Multimodal utterance-level affect analysis using visual, audio and text features. *arXiv preprint arXiv:1805.00625*.

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

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462.

Israel D Gebru, Sileye Ba, Xiaofei Li, and Radu Horaud. 2017. Audio-visual speaker diarization based on spatiotemporal bayesian fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1086–1099.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3454–3466, Brussels, Belgium. Association for Computational Linguistics.

Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111.

Zhen-Zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G Hauptmann. 2014. Multimedia classification and event detection using double fusion. *Multimedia tools and applications*, 71(1):333–347.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. 2019. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 2002–2006. IEEE.

Ziqian Luo, Hua Xu, and Feiyang Chen. 2019. Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network. In *AffCon@ AAAI*.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1899–1907. IEEE Computer Society.

Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4594–4602. IEEE Computer Society.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6892–6899. AAAI Press.

Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabás Poczós. 2018. Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 53–63, Melbourne, Australia. Association for Computational Linguistics.

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

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017b. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

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 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Interpretable multimodal routing for human multimodal language. *arXiv preprint arXiv:2004.14198*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2514–2520. ACM / IW3C2.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.

Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. 2016. Multilayer and multimodal fusion of deep neural networks for video classification. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 978–987.

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 Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5634–5641. AAAI Press.

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*.