Attention is not Enough: Mitigating the Distribution Discrepancy in Asynchronous Multimodal Sequence Fusion

Tao Liang[1,2] Guosheng Lin[3] Lei Feng[3] Yan Zhang[4] Fengmao Lv[1,5*]
[1] Southwest Jiaotong University
[2] Engineering Productivity & Quality Assurance of IES, Bytedance
[3] Nanyang Technological University
[4] University of Electronic Science and Technology of China
[5] Center of Statistical Research, Southwestern University of Finance and Economics
{fengmaolv,taoliangdpg}@126.com {gslin,feng0093}@ntu.edu.sg yixianqianzy@gmail.com

Abstract

Videos flow as the mixture of language, acoustic, and vision modalities. A thorough video understanding needs to fuse time-series data of different modalities for prediction. Due to the variable receiving frequency for sequences from each modality, there usually exists inherent asynchrony across the collected multimodal streams. Towards an efficient multimodal fusion from asynchronous multimodal streams, we need to model the correlations between elements from different modalities. The recent Multimodal Transformer (MulT) approach extends the self-attention mechanism of the original Transformer network to learn the crossmodal dependencies between elements. However, the direct replication of self-attention will suffer from the distribution mismatch across different modality features. As a result, the learnt crossmodal dependencies can be unreliable. Motivated by this observation, this work proposes the Modality-Invariant Crossmodal Attention (MICA) approach towards learning crossmodal interactions over modality-invariant space in which the distribution mismatch between different modalities is well bridged. To this end, both the marginal distribution and the elements with high-confidence correlations are aligned over the common space of the query and key vectors which are computed from different modalities. Experiments on three standard benchmarks of multimodal video understanding clearly validate the superiority of our approach.

1. Introduction

Videos analysis involves time-series data of language, acoustic, and vision modalities. Towards a thorough video understanding, we need to fuse the data sequences from different modalities. In practice, however, the collected multimodal streams are usually asynchronous due to the variable receiving frequency for sequences of different modalities [18]. For example, the sound or the subtitle may not exactly match what the video displays. The inherent asynchrony across different modalities raises a challenge on performing efficient multimodal fusion which requires to have precise information of the actual relationships between elements from different modality sequences.

To this end, the prior works manually preprocess the visual and acoustic sequences by aligning them to the resolution of textual words [15, 19, 24]. Then, multimodal fusion is performed on the word-aligned time steps. However, the manual alignment process usually requires a huge amount of time and labor effort. The recent Multimodal Transformer (MulT) approach extends the self-attention mechanism of the standard Transformer to learn the correlations between elements from different modalities [18]. Based on the latent crossmodal interaction explored via the crossmodal attention operations, MulT performs multimodal fusion directly from the asynchronous multimodal sequences without manual alignment.

However, if we take a further insight into the crossmodal attention mechanism in MulT, we will find that the direct replication of Transformer is suboptimal for asynchronous multimodal sequence fusion. In the standard Transformer model, the self-attention operations explore the correlations between elements by comparing the query and key vectors which are computed from the elements' features [21, 27]. On the other hand, MulT focuses on exploring the crossmodal correlations between elements. The query and key vectors involved in the crossmodal attention operation are computed from different modalities. Due to the heterogeneities across different modality features [9, 27], there

[*] Corresponding author: F. Lv (email: fengmaolv@126.com).

will exist a clear distribution mismatch in the common space of queries and keys. Hence, their dot-product cannot reveal reliable crossmodal correlations between elements.

Motivated by the above observation, this work proposes the Modality-Invariant Crossmodal Attention (MICA) approach towards multimodal fusion from asynchronous multimodal sequences. The core idea of our approach is to perform crossmodal attention over modality-invariant space in which the distribution discrepancy between different modalities is bridged. To this end, our approach enforces modality-invariance on the common space of the query and key vectors which are computed from different modalities. Overall, our approach bridges the distribution mismatch in two ways. One is to match the marginal distribution via Maximum Mean Discrepancy (MMD) which is commonly used in transfer learning or domain adaptation [13, 10]. The other is to match the elements with high-confidence correlations via our proposed Propagated Element-level Alignment (PEA) strategy. To be specific, our approach propagates the information of the crossmodal correlations along the network layers, i.e., the elements with high-confidence correlations in a previous layer will also participate in the element-level alignment loss (e.g., the L_2 loss) of the subsequent layers. The propagation strategy can enforce the consistency across the network layers and guide the Transformer network to progressively obtain better crossmodal correlations between elements. Compared with the original MulT model, our approach can overcome the distribution discrepancy between different modalities and build more reliable crossmodal relationships for multimodal fusion from asynchronous multimodal sequences. Experiments on three multimodal video understanding benchmarks clearly demonstrate the effectiveness of our approach.

To sum up, the contributions of this work are three-fold:

- We draw the first attention on the distribution discrepancy problem which restrains the attention mechanism to obtain reliable crossmodal correlations for asynchronous multimodal sequence fusion.
- We propose to perform crossmodal attention over modality-invariant space where the distribution gap across modalities is bridged. Both the marginal distribution mismatch and the element-level mismatch are aligned to reduce the distribution discrepancy.
- Our approach can obtain the state-of-the-art performance on different benchmarks of multimodal video understanding.

2. Related Works

Multimodal sequence fusion. A thorough video understanding needs to fuse data of different modalities, e.g., the language, acoustic, and vision modalities [26, 16, 17, 5, 19, 18]. The early works perform multimodal fusion on static features extracted from video clips and do not consider the inherent relationships between elements from different modality sequences [12, 5, 16, 17]. Towards an efficient multimodal fusion in videos, it is essential to take the inherent dependencies between elements of different modality sequences into consideration. However, the multimodal streams collected in practice are usually asynchronous due to the variable frame rate for sequences of different modalities [19, 22, 15]. To tackle this issue, the recent works manually preprocess the visual and acoustic sequences by aligning them to the resolution of textual words [19, 22, 15]. Based on the manual alignment, multimodal fusion is then implemented over the word-aligned elements. Typical works include hierarchical attention mechanism [8], nonverbal temporal interaction [22], cyclic translation [15], etc. However, the manual alignment process usually requires a huge amount of time and labor effort. Additionally, the word-level multimodal fusion does not consider the long term dependencies between elements across modalities.

Zeng et al. propose to perform multimodal fusion directly from the unaligned multimodal sequences via the maximum mutual information rule [25]. The performance of their approach is heavily limited by the shallow architecture. Recently, Tsail et al. propose to extend the Transformer network to learn the latent correlations between elements of different modalities [18]. The recent work improves MulT by introducing a common message hub to reinforcing each modality [11]. However, a lot of additional parameters are required in [11].

Distribution alignment. Distribution alignment is originally studied for domain adaptation [13, 10, 6, 20, 3]. The common distribution alignment approaches include maximum mean discrepancy [13, 10], adversarial training [6, 20], adaptive batch normalization [3], etc. This work draws an interesting insight connecting Transformer with domain adaptation via distribution alignment. We note that our work has a different motivation for distribution alignment. In particular, domain adaptation bridges the distribution mismatch to improve the model's generalization ability across different domains [13]. On the other hand, this work mainly focuses on modeling more reliable correlations between elements across modalities.

3. Modality-Invariant Crossmodal Attention

3.1. Problem statement

This work focuses on performing multimodal sequence fusion from the three major modalities in videos, i.e., the language (L), vision (V), and acoustic (A) modalities. With the sequence length and the feature dimension denoted by $T_{(.)}$ and $d_{(.)}$, respectively, we use the notations $X_{\{L,V,A\}} \in$

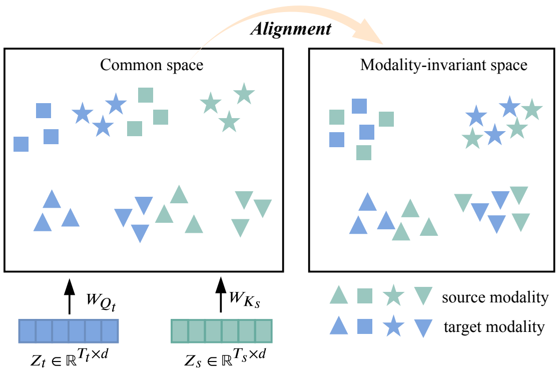


Figure 1. The distribution discrepancy between the queries and keys in the attention mechanism caused by the heterogeneities across different modality features. The elements with inherent correlations are displayed in the same shape.

$\mathbb{R}^{T_{\{L,V,A\}} \times d_{\{L,V,A\}}}$ to represent the input sequences from each modality. Due to the variable receiving frequency for sequences from each modality, there usually exists inherent asynchrony across different multimodal sequences. In this work, our goal is to perform multimodal fusion from the asynchronous multimodal sequences and obtain representations which are effective for downstream prediction tasks such as human emotion recognition.

3.2. Preliminaries

Crossmodal attention. The crossmodal attention operation is the core component of the MulT model [18]. It receives inputs from a source modality and a target modality and focuses on modeling the correlations between elements across modalities. With $s, t \in \{L, V, A\}$, we use the notations $X_s \in \mathbb{R}^{T_s \times d_s}$ and $X_t \in \mathbb{R}^{T_t \times d_t}$ to represent the data sequences from the source and target modalities, respectively. Similar to the self-attention mechanism in Transformer, the crossmodal attention operation also involves the queries, keys, and values, which are represented as $Q_t = X_t W_{Q_t}$, $K_s = X_s W_{K_s}$ and $V_s = X_s W_{V_s}$, respectively. The weights $W_{Q_t} \in \mathbb{R}^{d_t \times d_k}$, $W_{K_s} \in \mathbb{R}^{d_s \times d_k}$ and $W_{V_s} \in \mathbb{R}^{d_s \times d_v}$ are learnable parameters. One single head of the crossmodal attention operation can be formulated as follows:

$$\begin{aligned} Y_t &= \mathrm{CM}_{s \to t}(X_s, X_t) \\ &= \mathrm{softmax}(\frac{Q_t K_s^T}{\sqrt{d_k}})V_s \\ &= \mathrm{softmax}(\frac{X_t W_{Q_t} W_{K_s}^T X_s^T}{\sqrt{d_k}})X_s W_{V_s}, \end{aligned} \tag{1}$$

where $Y_t \in \mathbb{R}^{T_t \times d_v}$. We denote the whole h-head crossmodal attention operation as $Y_t = \mathrm{CM}^{\mathrm{mul}}_{s \to t}(X_s, X_t)$, where $Y_t \in \mathbb{R}^{T_t \times h d_v}$. As shown in Eq. 1, the multimodal fusion occurs by attending to the correlated elements of the source modality. Y_t will be used to reinforce the target modality features. We refer the readers to [18] for more details.

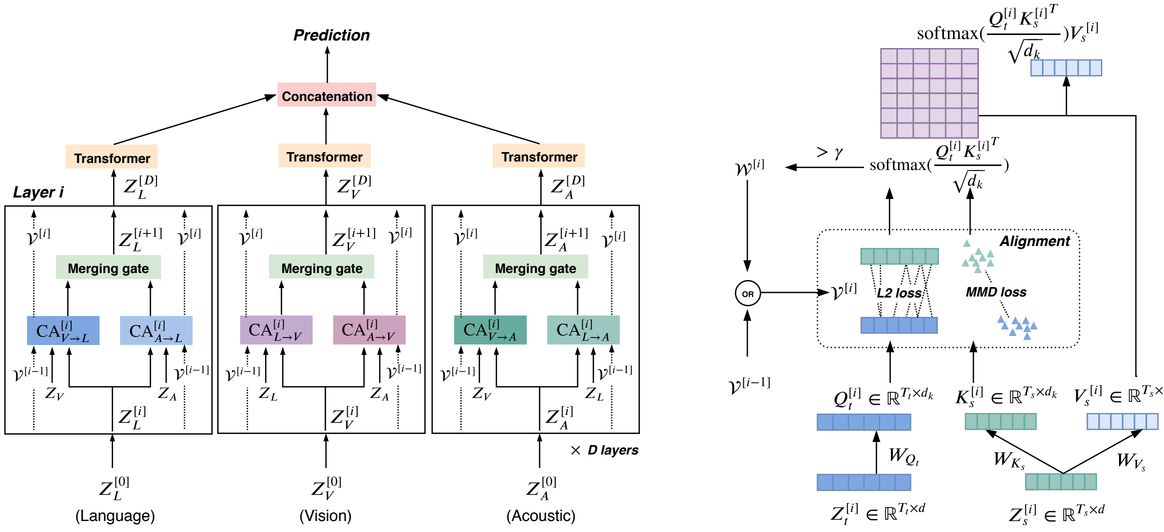
3.3. Motivation

In the crossmodal attention operation, W_{K_s} and W_{Q_t} first project the elements of the source and target modalities into the common space as $K_s \in \mathbb{R}^{T_s \times d_k}$ and $Q_t \in \mathbb{R}^{T_t \times d_k}$, respectively (see Fig. 1). The crossmodal correlations between elements are then explored by comparing K_s and Q_t in the common space. However, unlike self-attention in the standard Transformer, the queries and keys herein are computed from different modalities. Due to the heterogeneities across different modality features, there will exist a clear distribution mismatch between K_s and Q_t, i.e., $\mathcal{P}(K_s) \neq \mathcal{P}(Q_t)$. Similar to the troubles in domain adaptation, the distribution mismatch will make the crossmodal correlations observed from the common space unreliable. For example, two elements which should be related may have a large distance over the common space or vice versa due to the distribution mismatch between the queries and keys (see Fig. 1). Motivated by this observation, we wonder whether better crossmodal correlations can be modeled by aligning the distribution discrepancy across modalities and propose the MICA approach towards multimodal fusion from asynchronous multimodal sequences.

3.4. Modality-invariant crossmodal attention

Network backbone. As in [18], the original sequences are first preprocessed by a 1D temporal convolutional layer and a positional embedding augment opertaion. The features of different modalities are enforced to have the same dimension by controlling the kernel size of the 1D convolutional operation used for each modality. We use the notations $Z_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d}$ to represent the preprocessed sequences. $Z_{\{L,V,A\}} \in \mathbb{R}^{T_{\{L,V,A\}} \times d}$ will be used as the inputs of the Transformer network. Fig. 2 displays the overall architecture of the proposed approach. The backbone is almost similar to the MulT model. In the network, multiple stacks of pairwise and bidirectional crossmodal attention blocks are used to update the sequences. Each crossmodal attention block reinforces a target modality by attending to the correlated elements of a source modality based on the attention mechanism (see Fig. 3(b) of [18]). We then pass the reinforced sequences of each modality through a self-attention transformer and concatenate them as the representation for downstream prediction tasks. Several fully-connected layers are used to make the final prediction. The model is trained in an end-to-end manner.

We note that our network backbone has a small difference with the MulT model. Denote by $Z^{[i]}_{s_1 \to t} \in \mathbb{R}^{T_t \times d}$ and $Z^{[i]}_{s_2 \to t} \in \mathbb{R}^{T_t \times d}$ the modalities of t reinforced by the modalities of s_1 and s_2, respectively, where $s_1, s_2, t \in \{L, V, A\}$. The superscript $[i]$ represents the i-th layer. After being



(a) The overall architecture of the proposed approach.

(b) The distribution alignment operation in crossmodal attention blocks.

Figure 2. (a) The overall architecture of the proposed MICA approach. The dotted line represents propagating the crossmodal correlations across the network layers via the weight matrix $\mathcal{V}^{[i]}$. The notation $\mathrm{CA}_{s\to t}^{[i]}$, where $s, t \in \{L, V, A\}$, denotes the crossmodal attention blocks. (b) The distribution alignment operation in the crossmodal attention block $\mathrm{CA}_{s\to t}^{[i]}$. The distribution mismatch is aligned over the common space of queries and keys computed from different modalities.

reinforced by the crossmodal attention block, $Z_{s_1\to t}^{[i]}$ and $Z_{s_2\to t}^{[i]}$ are merged via the following gate:

$$G_t^{[i]} = \mathrm{sigmoid}(Z_{s_1\to t}^{[i]} \cdot W_{s_1\to t}^{[i]} + Z_{s_2\to t}^{[i]} \cdot W_{s_2\to t}^{[i]} + b_t^{[i]}),$$

$$Z_t^{[i+1]} = G_t^{[i]} \odot Z_{s_1\to t}^{[i]} + (1 - G_t^{[i]}) \odot Z_{s_2\to t}^{[i]},$$

where $W_{s_1\to t}^{[i]} \in \mathbb{R}^{d\times d}$, $W_{s_2\to t}^{[i]} \in \mathbb{R}^{d\times d}$, and $b_t^{[i]} \in \mathbb{R}^{T_t\times d}$ are learnable parameters. $Z_{s_1\to t}^{[i]}$ and $Z_{s_2\to t}^{[i]}$ are merged with different proportions determined via the learnable parameters. $Z_t^{[i+1]}$ will be input into the crossmodal attention blocks of the next layer. Unlike the MulT model in which $Z_{s_1\to t}^{[i]}$ and $Z_{s_2\to t}^{[i]}$ flow separately and are concatenated together at the final stage, the merge of $Z_{s_1\to t}^{[i]}$ and $Z_{s_2\to t}^{[i]}$ at intermediate layers can promote the effective cooperation between the crossmodal attention blocks which focus on the same target modality (e.g., $\mathrm{CA}_{V\to L}^{[i]}$ and $\mathrm{CA}_{A\to L}^{[i]}$ in Fig. 2) and make the multimodal fusion more efficient.

Model overview. As discussed in Section 3.3, the crossmodal attention will suffer from the distribution discrepancy across different modality features. Hence, our approach mainly focuses on aligning the distribution discrepancy across modalities. To this end, we enforce modality-invariance on the common space of the queries and keys which are computed from different modalities. Crossmodal attention is then performed over the modality-invariant space in which the distribution mismatch between different modalities is already bridged.

In general, the distribution discrepancy across modalities lies in two aspects. One is the mismatch of the marginal distribution. The other is the element-level mismatch, e.g., two elements with actual correlations may be projected far away from each other. Hence, our approach bridges the distribution discrepancy in two ways. For the former, we reduce the MMD metrics between the queries and keys in each crossmodal attention unit. The element-level mismatch is aligned via our proposed PEA approach. Denote by $\mathcal{L}_p$ the cross-entropy loss of the downstream prediction task, $\mathcal{L}_m$ the alignment loss for the marginal distribution, and $\mathcal{L}_e$ the element-level alignment loss. The overall objective can be represented as follows:

$$\mathcal{L} = \mathcal{L}_p + \alpha\mathcal{L}_m + \beta\mathcal{L}_e,$$

where α and β are the trade-off parameters that weigh the importance of the corresponding terms. $\mathcal{L}_m$ and $\mathcal{L}_e$ are formulated as follows:

Marginal distribution alignment. With $s, t \in \{L, A, V\}$, we represent the element representations of the source and target modalities involved in the crossmodal attention operation $\mathrm{CM}_{s\to t}^{[i]}$ as $\boldsymbol{z}_s^{[i]} \in \mathbb{R}^{d\times 1}$ and $\boldsymbol{z}_t^{[i]} \in \mathbb{R}^{d\times 1}$, respectively. $\mathrm{CM}_{s\to t}^{[i]}$ will project $\boldsymbol{z}_s^{[i]}$ and $\boldsymbol{z}_t^{[i]}$ into a common space as keys and queries, respectively:

$$\boldsymbol{k}_s^{[i]} = W_{K_s}^{[i]\,T} \boldsymbol{z}_s^{[i]}, \ \boldsymbol{q}_t^{[i]} = W_{Q_t}^{[i]\,T} \boldsymbol{z}_t^{[i]},$$

where $W_{K_s}^{[i]}, W_{Q_t}^{[i]} \in \mathbb{R}^{d\times d_k}$. The marginal distribution mismatch is bridged by performing the MMD alignment over the common space of $\boldsymbol{k}_s^{[i]}$ and $\boldsymbol{q}_t^{[i]}$. To this end, we map $\boldsymbol{k}_s^{[i]}$ and $\boldsymbol{q}_t^{[i]}$ into a Reproducing Kernel Hilbert Space (RKHS) endowed with a characteristic kernel and measure the distribution difference via two-sample test:

$$\ell_{s,t}^{[i]} = \left\| \mathbb{E}_{\boldsymbol{z}_s^{[i]}}[\phi(\boldsymbol{k}_s^{[i]})] - \mathbb{E}_{\boldsymbol{z}_t^{[i]}}[\phi(\boldsymbol{q}_t^{[i]})] \right\|_{\mathcal{H}}^2 ,$$

where $\phi(.)$ denotes the mapping to RKHS $\mathcal{H}$. Denote by n_s and n_t the total number of the elements in the source and target modalities, respectively. The empirical estimation of $\ell_{s,t}^{[i]}$ is computed by

$$\begin{aligned}\hat{\ell}_{s,t}^{[i]} =& \frac{1}{n_s^2} \sum_{m=1}^{n_s} \sum_{n=1}^{n_s} K(\boldsymbol{k}_{s,m}^{[i]}, \boldsymbol{q}_{s,n}^{[i]}) + \frac{1}{n_t^2} \sum_{m=1}^{n_t} \sum_{n=1}^{n_t} K(\boldsymbol{k}_{t,m}^{[i]}, \boldsymbol{q}_{t,n}^{[i]}) \\ &- \frac{2}{n_s n_t} \sum_{m=1}^{n_s} \sum_{n=1}^{n_t} K(\boldsymbol{k}_{s,m}^{[i]}, \boldsymbol{q}_{t,n}^{[i]}),\end{aligned}$$

where $K(. \ , \ .)$ represents the kernel function. In our approach, the MMD alignment is performed in the crossmodal attention blocks of each layer. Denote by D the number of Transformer layers. $\mathcal{L}_m$ can be formulated as follows:

$$\mathcal{L}_m = \sum_{i=0}^{D} \sum_{s,t \in \{L,V,A\}, s \neq t} \hat{\ell}_{s,t}^{[i]} .$$

Propagated element-level alignment. The element-level alignment mainly focuses on bridging the distribution mismatch of the elements with actual correlations. However, the actual crossmodal correlations between elements are unknown for the asynchronous multimodal sequences. To address this problem, we leverage the information revealed from the crossmodal attention operations, i.e., $\mathrm{softmax}(\frac{Q_t K_s^T}{\sqrt{d_k}})$ in Eq. 1. Specifically, this matrix estimates the probability that two elements from the sequences of different modalities have an actual correlation. If two elements have a large probability of being correlated, we can assume that there exists an actual corresponding relation between them. Immediately, we reduce the L_2 distance between the corresponding query and key vectors:

$$d_{j,s,t}^{[i]} = \sum_{n=1}^{T_t} \sum_{m=1}^{T_s} \mathcal{W}_{n,m}^{[i]} \cdot \left\| \boldsymbol{q}_{j,t,n}^{[i]} - \boldsymbol{k}_{j,s,m}^{[i]} \right\|^2 ,$$

with the weight $\mathcal{W}_{n,m}^{[i]}$ computed by

$$\mathcal{W}^{[i]} := \mathrm{softmax}(\frac{Q_t^{[i]} {K_s^{[i]}}^T}{\sqrt{d_k}}) > \gamma,$$

where γ is the threshold dynamically set to the probability ranked at $\tau * T_s * T_t$ in the matrix $\mathrm{softmax}(\frac{Q_t K_s^T}{\sqrt{d_k}})$. The selection rate $\tau \in [0, 1]$ is a pre-defined hyper-parameter. We note that $\boldsymbol{k}_{j,s,m}^{[i]}$ and $\boldsymbol{q}_{j,t,n}^{[i]}$ are from the sequences of the same training sample (denoted by the subscript j).

This shares a similar idea of the common self-learning approach [28, 29]. Beyond self-learning, however, our approach further draws attentions on the inconsistency cross the network layers. In practice, the crossmodal attention operations of different layers can model different correlations between elements. For example, the crossmodal correlations modeled in a previous layer may not be observed in the subsequent layers. As a result, the element-level alignment can be inconsistent across the network layers. To address this issue, our approach propagates the information of the crossmodal correlations across the network layers via the weight matrix $\mathcal{V}^{[i]}$: $\mathcal{V}^{[l]} = \bigcup_{i=0}^{l} \mathcal{W}^{[i]}$. The element-level alignment loss is then weighed by $\mathcal{V}_{n,m}^{[l]}$ instead of $\mathcal{W}_{n,m}^{[i]}$:

$$d_{j,s,t}^{[i]} = \sum_{n=1}^{T_t} \sum_{m=1}^{T_s} \mathcal{V}_{n,m}^{[i]} \cdot \left\| \boldsymbol{q}_{j,t,n}^{[i]} - \boldsymbol{k}_{j,s,m}^{[i]} \right\|^2 . \qquad (2)$$

Note that $\mathcal{V}^{[0]}$ is initialized as $\mathcal{W}^{[0]}$. Weighed by $\mathcal{V}_{n,m}^{[l]}$, the elements with high-confidence correlations in a previous layer will also participate in the element-level alignment loss of the subsequent layers. This strategy can enforce consistency across the network layers and guide the Transformer model to progressively model more reliable crossmodal correlations between elements. With $d_{j,s,t}^{[i]}$ defined in Eq. 2, the $\mathcal{L}_e$ loss can be formulated as follows:

$$\mathcal{L}_e = \sum_{j=1}^{N} \sum_{i=0}^{D} \sum_{s,t \in \{L,V,A\}, s \neq t} d_{j,s,t}^{[i]},$$

where N represents the number of training samples.

4. Experiments

4.1. Experimental setup

We conduct experiments on three standard benchmarks of multimodal video understanding, including CMU-MOSI [24], CMU-MOSEI [23] and IEMOCAP [2]. These benchmarks mainly focus on human multimodal emotion recognition which requires to perform an efficient multimodal sequence fusion. The common protocol of the previous works [18, 19, 22] is adopted in our experiments.

CMU-MOSI is a dataset consisting of 2,199 samples of short monologue video clips [24]. Its predetermined data partition has 1,284 samples in the training set, 229 in the validation set, and 686 in the testing set. Each sample is labeled with a sentiment score ranging from -3 (very negative) to 3 (very positive). The acoustic and visual sequences are extracted at the receiving frequency of 12.5 and 15 Hz, respectively. As in the previous works [18, 19], the performance is evaluated by the 7-class accuracy (i.e., Acc_7), binary accuracy (i.e., Acc_2) and F1 score.

Table 1. The hyperparameter settings adopted in each benchmark.

Setting	CMU-MOSEI	CMU-MOSI	IEMOCAP
Optimizer	Adam	Adam	Adam
Batch size	64	64	32
Epoch number	120	120	80
Learning rate	5e-4	1e-3	1e-3
Feature size d	40	40	40
Attention head h	10	8	10
Selection rate τ	0.3	0.25	0.25
tradeoff parameter α	0.8	0.8	0.7
tradeoff parameter β	0.5	0.5	0.5
Kernel size (L/V/A)	3/3/3	3/3/3	3/3/5
Transformer layer D	6	4	4

Table 2. Comparison on the CMU-MOSI benchmark. The superscript † indicates that a manual alignment process is needed.

Method	Acc$_7$(%)	Acc$_2$(%)	F1(%)
EF-LSTM	31.0	73.6	74.5
LF-LSTM	33.7	77.6	77.8
MFM† [19]	36.2	78.1	78.1
RAVEN [22]	31.7	72.7	73.1
MCTN [15]	32.7	75.9	76.4
MulT [18]	39.1	81.1	81.0
MICA (ours)	**40.8**	**82.6**	**82.7**

CMU-MOSEI is a dataset made up of 22,856 samples of movie review video clips [23]. Its predetermined data partition has 16,326 samples in the training set, 1,871 in the validation set, and 4,659 in the testing set. As in the above setting, the CMU-MOSEI samples are also labeled with the sentiment scores ranging from -3 to 3. The acoustic and visual sequences are extracted at the receiving frequency of 20 and 15 Hz, respectively. The performance metrics are the same to the ones used in the above setting.

IEMOCAP is a dataset consisting of 4,453 samples of video clips [2]. Its predetermined data partition has 2,717 samples in the training set, 798 in the validation set, and 938 in the testing set. The acoustic and visual sequences are extracted at the receiving frequency of 12.5 and 15 Hz, respectively. Different from CMU-MOSI and CMU-MOSEI, this benchmark mainly focuses on multi-label learning [22]. The models are required to recognize 4 emotion classes (i.e., happy, sad, angry and neutral) from video clips. As in the previous works [19, 22], the performance is evaluated by the binary classification accuracy and the F1 score for each emotion class.

4.2. Implementation details

To extract features of the visual modality, the Facet model is used to preprocess the video frames [1]. For each video frame, 35 facial action units are generated to represent the facial muscle movement. For the textual modality, the pre-trained Glove model is used to convert the video transcripts [14]. Each textual word is represented by a 300-dimensional word embedding. The COVAREP model is used to extract the acoustic features [4]. The dimension of the acoustic features is 74.

Table 3. Comparison on the CMU-MOSEI benchmark. The superscript † indicates that a manual alignment process is needed.

Method	Acc$_7$(%)	Acc$_2$(%)	F1(%)
EF-LSTM	46.3	76.1	75.9
LF-LSTM	48.8	77.5	78.2
GMFN† [24]	45.0	76.9	77.0
RAVEN [22]	45.5	75.4	75.7
MCTN [15]	48.2	79.3	79.7
MulT [18]	50.7	81.6	81.6
MICA (ours)	**52.4**	**83.7**	**83.3**

The hyperparameters adopted in each benchmark are displayed in Table 1. The kernel size is set for the 1D temporal convolutional operation used to preprocess the input sequence of each modality. The hyper-parameters are determined via the validation set.

4.3. Experimental Results

Baselines. We compare our approach with the recent state-of-the-art works for multimodal sequence fusion, including Early Fusion LSTM (EF-LSTM), Late Fusion LSTM (LF-LSTM), Multimodal Factorization Model (MFM) [19], Graph Multimodal Fusion Network (GMFN) [24], Recurrent Attended Variation Embedding Network (RAVEN) [22], Multimodal Cyclic Translation Network (MCTN) [15] and Multimodal Transformer (MulT) [18]. Of these, MFM and G-MFN require a manual process to align the asynchrony across modalities. RAVEN and MCTN can be applicable for multimodal fusion from asynchronous multimodal sequences by including an additional Connectionist Temporal Classification (CTC) loss [7] into the their learning objectives. MulT and LF-LSTM are directly applicable for asynchronous multimodal sequences.

Performance comparison. We report the experimental results of each baseline in Table 2 - 4. In general, three observations can be drawn as follows. First, we can obtain the state-of-the-art results on the adopted benchmarks which involve multimodal fusion from asynchronous multimodal sequences. Second, our approach outperforms MFM and GMFN without manually aligning the asynchronous multimodal sequences. Finally, our approach has a clear performance improvement compared with the MulT model. Compared with MulT, the improvement of our approach is significant for all the metrics in each benchmark ($p < 0.05$).

For a more fair comparison between our approach and MulT, we also conduct experiments by controlling the num-

Table 4. Comparison on the IEMOCAP benchmark in terms of the binary classification accuracy and the F1 score for each emotion class.

Method	Happy		Sad		Angry		Neutral	
	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4
LF-LSTM	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2
RAVEN [22]	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5
MCTN [15]	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3
MulT [18]	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7
MICA (ours)	**86.8**	**83.9**	**79.3**	**75.2**	**75.7**	**72.4**	**63.7**	**61.6**

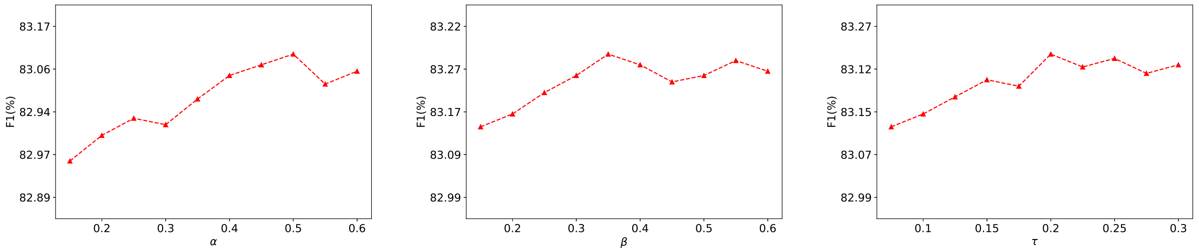


Figure 3. Sensitivity analysis on the CMU-MOSEI benchmark. The results are obtained by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments.

Table 5. Ablation study results on the CMU-MOSEI benchmark. The notations "MMD align" and "PE align" denote the marginal distribution alignment and propagated element-level alignment, respectively. The results are averaged over 5 runs.

Model design	Acc_7(%)	Acc_2(%)	F1(%)
Backbone w/o align	51.1±0.14	81.9±0.26	81.9±0.21
MMD align	52.0±0.11	83.2±0.17	82.9±0.21
MMD align + PE align (full model)	52.4±0.03	83.7±0.12	83.3±0.13
MMD align + PE align w/o propagation	52.2±0.04	83.4±0.13	83.0±0.16

Table 6. Analysis of cross-modal alignment on CMU-MOSEI. The $\mathcal{A}$-distance is estimated between the queries and keys from the last transformer layer. The notation "Corr. Acc" denotes the accuracy of the modeled cross-modal element-level correlations. The results are obtained from the testing data.

Model design	$\mathcal{A}$-distance	Corr. Acc(%)
MulT w/o align	1.63	80.4
Backbone w/o align	1.69	81.8
MICA	1.37	86.7

ber of epochs and transformer layers. The results further demonstrate the effectiveness of our approach (see Table A1 of the supplementary).

4.4. Analysis

Ablation study. We conduct the the ablation study on the CMU-MOSEI benchmark and report the results in Table 5. The first row displays the performance of the backbone model. We can see that our backbone network can obtain better performance than the original MulT model. This observation supports the effectiveness of the merging gate incorporated in each layer of the backbone network. In the next two rows, the marginal distribution alignment loss and the propagated element-level alignment loss are gradually included into the model. It is clear that both of them improve the performance effectively. This observation clearly demonstrates the necessity of the distribution alignment proposal. With the distribution mismatch between different modalities well bridged, the attention mechanism can be more suitable for modeling the crossmodal correlations between elements. Furthermore, we remove the propagation mechanism from the element-level alignment loss and implement Eq. 2 by replacing $\mathcal{V}^{[i]}_{n,m}$ with $\mathcal{W}^{[i]}_{n,m}$. As reported in the last row, the performance improvement of the standard element-level alignment is limited. Without the propagation mechanism, the element-level distribution alignment will be inconsistent across different network layers. The propagation strategy can enforce consistency across the network layers and guide the Transformer network to progressively learn more reliable crossmodal correlations.

We also provide the ablation study results on the standard MulT backbone in the supplementary. The alignment losses are also effective on the original MulT (see Table A2).

Distribution discrepancy. We further conduct the analysis on the problem of distribution discrepancy pointed in this paper. From Table 6, we can see that the proposed alignment losses can help to reduce the $\mathcal{A}$-distance (i.e., a common measure of domain discrepancy) between queries and keys computed from different modalities, as well as model

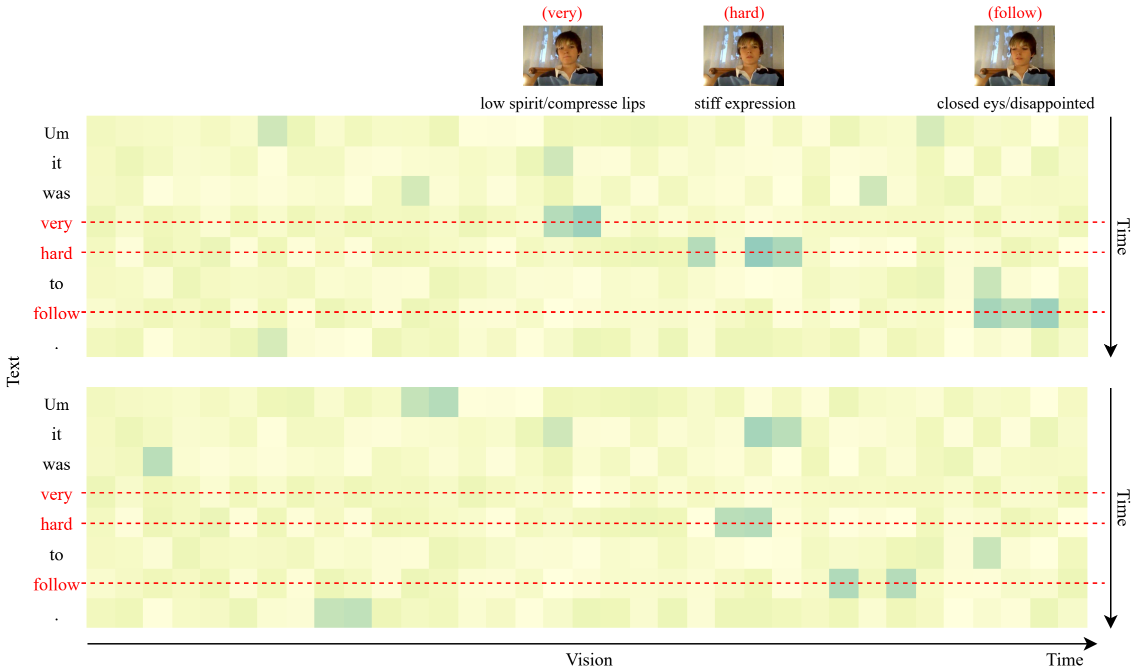


Figure 4. Visualization analysis of the modeled crossmodal dependencies between elements in the CMU-MOSI benchmark. The visualization cases of the full MICA approach and the non-alignment backbone are shown in the upper part and the bottom part, respectively. The textual words displayed above each video frame are the corresponding spoken words. The results are from the crossmodal attention unit of the fourth Transformer layer.

significantly better cross-modal correlations between elements (the manual element-level alignment is used as the ground truth). Without the alignment losses, the modeled cross-modal correlations are much worse. This observation supports that the alignment losses improve the performance by modeling better cross-modal correlations.

Sensitivity analysis. Moreover, the sensitivity analysis for hyper-parameters is conducted on CMU-MOSEI, in order to verify the robustness of our approach. The tested hyper-parameters include the trade-off parameters α for the $\mathcal{L}_m$ loss, the trade-off parameters β for the $\mathcal{L}_e$ loss, and the selection rate τ in element-level distribution alignment. In particular, the sensitivity analysis is conducted by varying the value of the corresponding hyper-parameter, while fixing the other hyper-parameters to the values adopted in the experiments. We display the sensitivity analysis results in Fig.3. It is clear that the performance of the proposed approach is not sensitive to the values of the hyper-parameters.

Qualitative analysis. Finally, we display the visualization examples of the modeled crossmodal dependencies between elements in the CMU-MOSI benchmark. From Fig. 4, we can see that our approach models a reasonable correlation between the video frames and the spoken words. The emotion related words successfully attend to the video frames which contains the corresponding facial expression. On the other hand, the crossmodal correlations modeled in the non-alignment backbone are meaningless.

5. Conclusion

This work proposes the modality-invariant crossmodal attention approach towards learning the crossmodal interactions between elements from asynchronous multimodal sequences in videos. Our approach draws attentions on the distribution shift problem in Transformer caused by the heterogeneities across different modalities. To model better crossmodal correlations, we propose to perform crossmodal attention over modality-invariant space where the distribution shift across modalities is bridged. Both the marginal distribution mismatch and the element-level mismatch are considered. Experiments on different benchmarks clearly support the superiority of our approach.

Acknowledgements. This work was conducted when T. Liang worked as a research assistant at Southwest Jiaotong University, supervised by F. Lv. F. Lv and T. Liang contributed equally to this work. This work was supported by the National Natural Science Foundation of China (No.6210 6204) and the Fundamental Research Funds for the Central Universities of China (No.JBK1806002). G. Lin's participation was supported by an NTU Start-up Grant and MOE Tier-1 research grants: RG28/18 (S), RG22/19 (S) and RG95/20.

References

[1] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *WACV*, pages 1–10, 2016.

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

[3] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, pages 7354–7362, 2019.

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP - A collaborative voice analysis repository for speech technologies. In *ICASSP*, pages 960–964, 2014.

[5] Quan Gan, Shangfei Wang, Longfei Hao, and Qiang Ji. A multimodal deep regression bayesian network for affective video content analyses. In *ICCV*, pages 5123–5132, 2017.

[6] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, volume 37, pages 1180–1189, 2015.

[7] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In William W. Cohen and Andrew W. Moore, editors, *ICML*, volume 148, pages 369–376, 2006.

[8] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *ACL*, pages 2225–2235, 2018.

[9] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. Heterogeneous domain adaptation through progressive alignment. *IEEE Trans. Neural Networks Learn. Syst.*, 30(5):1381–1391, 2019.

[10] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, volume 37, pages 97–105, 2015.

[11] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *CVPR*, pages 2554–2562, 2021.

[12] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[13] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22(2):199–210, 2011.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[15] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, pages 6892–6899, 2019.

[16] Dung Nguyen Tien, Kien Nguyen, Sridha Sridharan, David Dean, and Clinton Fookes. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Comput. Vis. Image Underst.*, 174:33–42, 2018.

[17] Dung Nguyen Tien, Kien Nguyen Thanh, Sridha Sridharan, Afsane Ghasemi, David Dean, and Clinton Fookes. Deep spatio-temporal features for multimodal emotion recognition. In *WACV*, pages 1215–1223, 2017.

[18] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, pages 6558–6569, 2019.

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

[20] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 2962–2971, 2017.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[22] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, pages 7216–7223, 2019.

[23] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*, pages 2236–2246, 2018.

[24] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.*, 31(6):82–88, 2016.

[25] Zhihong Zeng, Jilin Tu, Brian Pianfetti, Ming Liu, Tong Zhang, ZhenQiu Zhang, Thomas S. Huang, and Stephen E. Levinson. Audio-visual affect recognition through multi-stream fused HMM for HCI. In *CVPR*, pages 967–972, 2005.

[26] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur A. Ciftci, Shaun J. Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPP*, pages 3438–3446, 2016.

[27] Joey Tianyi Zhou, Ivor W. Tsang, Sinno Jialin Pan, and Mingkui Tan. Multi-class heterogeneous domain adaptation. *J. Mach. Learn. Res.*, 20:57:1–57:31, 2019.

[28] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, volume 11207, pages 297–313, 2018.

[29] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, pages 5981–5990, 2019.