Information Fusion 102 (2024) 102031

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Disentanglement Translation Network for multimodal sentiment analysis

Ying Zeng, Wenjun Yan, Sijie Mai, Haifeng Hu *

School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510006, China

ARTICLE INFO

Keywords:
Mutlimodal representation learning
Disentanglement learning
Multimodal sentiment analysis
Feature reconstruction

ABSTRACT

Obtaining an effective joint representation has always been the goal for multimodal tasks. However, distributional gap inevitably exists due to the heterogeneous nature of different modalities, which poses burden on the fusion process and the learning of multimodal representation. The imbalance of modality dominance further aggravates this problem, where inferior modalities may contain much redundancy that introduces additional variations. To address the aforementioned issues, we propose a Disentanglement Translation Network (DTN) with Slack Reconstruction to capture desirable information properties, obtain a unified feature distribution and reduce redundancy. Specifically, the encoder–decoder-based disentanglement framework is adopted to decouple the unimodal representations into modality-common and modality-specific subspaces, which explores the cross-modal commonality and diversity, respectively. In the encoding stage, to narrow down the discrepancy, a two-stage translation is devised to incorporate with the disentanglement learning framework. The first stage targets at learning modality-invariant embedding for modality-common information with adversarial learning strategy, capturing the commonality shared across modalities. The second stage considers the modality-specific information that reveals diversity. To relieve the burden of multimodal fusion, we realize Specific-Common Distribution Matching to further unify the distribution of the desirable information. As for the decoding and reconstruction stage, we propose Slack Reconstruction to seek a balance between retaining discriminative information and reducing redundancy. Although the existing commonly-used reconstruction loss with strict constraint lowers the risk of information loss, it easily leads to the preservation of information redundancy. In contrast, Slack Reconstruction imposes a more relaxed constraint so that the redundancy is not forced to be retained, and simultaneously explores the inter-sample relationships. The proposed method aids multimodal fusion by learning the exact properties and obtaining a more uniform distribution for cross-modal data, and manages to reduce information redundancy to further ensure feature effectiveness. Extensive experiments on the task of multimodal sentiment analysis indicate the effectiveness of the proposed method. The codes are available at https://github.com/zengy268/DTN.

1. Introduction

The development of smart devices enables the availability of abundant multimodal data, which facilitates various application including multimodal sentiment analysis, multimodal emotion recognition, visual question answering, etc. Multimodal sentiment analysis is one of the most important topics that aim to understand the implicit sentiment from cross-modal input (i.e., language, audio, and visual). Previous researchers endeavor to design various methods to mine the latent information behind the collected data [1–10], and most of them rely on sophisticated fusion methods to explore the intra-modal and inter-modal dynamics. Despite the effectiveness of the fusion methods, the existence of modality gap still poses great challenge to multimodal models. Although different modalities provide complementary information and distinct characteristics from different aspects that are beneficial to learning a more comprehensive and effective joint embedding, their heterogeneous nature inevitably causes distributional discrepancy in the feature space. The distributional gap between modalities leads to the difficulty in mining complementary information across modalities [11], and negatively affects the fusion process to obtain effective multimodal representation. The imbalance of modality dominance even aggravates this problem. In many multimodal learning tasks, it is not always the case that all modalities contain sufficient information and show high unimodal precision. For example, in the field of MSA, the non-lexical modalities are inferior than the language one, and they may contain much redundant information. This useless information

* Corresponding author.
E-mail addresses: zengy268@mail2.sysu.edu.cn (Y. Zeng), yanwj23@mail2.sysu.edu.cn (W. Yan), maisj@mail2.sysu.edu.cn (S. Mai), huhaif@mail.sysu.edu.cn (H. Hu).

https://doi.org/10.1016/j.inffus.2023.102031
Received 30 June 2023; Received in revised form 15 September 2023; Accepted 18 September 2023
Available online 21 September 2023

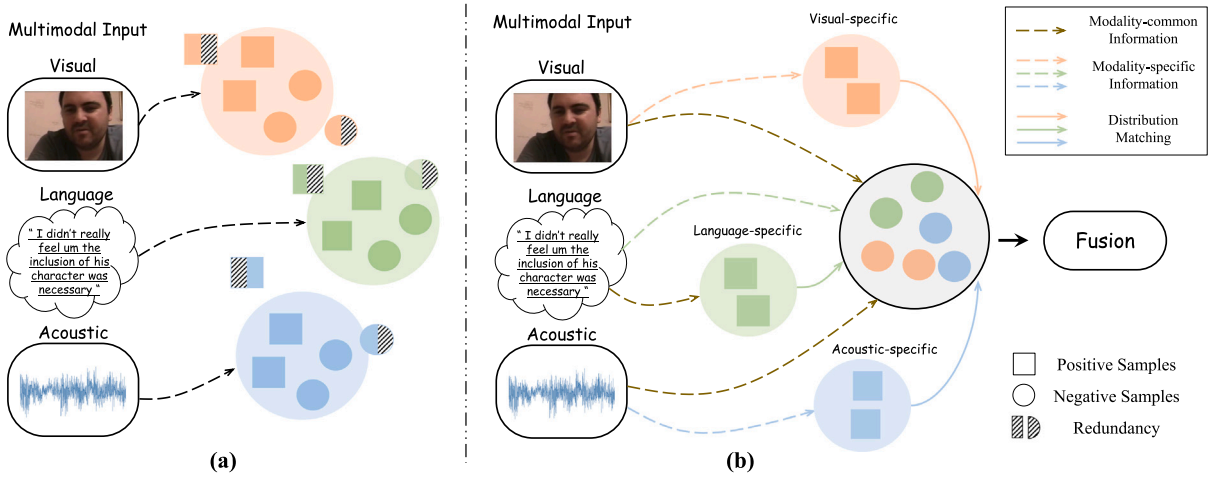

Fig. 1. (a) Distributional discrepancy between modalities poses burden on multimodal fusion and inference. Redundant information further leads to variations in feature space. (b) The proposed method decouples individual modalities into desirable properties to aid fusion, and reduces variations brought by useless or noisy information.

brings variations in the feature space, leading to further gap between modalities and affect multimodal fusion.

To address the aforementioned issues, we propose a novel learning framework Disentanglement Translation Network with Slack Reconstruction to achieve a more uniform distribution for cross-modal data. As shown in Fig. 1, the proposed method learns desirable properties of the multimodal input, narrows down the distributional discrepancy, and meanwhile reduces redundant information contained in the original input. Specifically, the proposed method is based on the encoder–decoder framework. In the feature encoding stage, we decouple the original feature into two desirable information properties (i.e., modality-common and -specific information). To achieve a uniform distribution regardless of modalities and information properties, Disentanglement-based Translation is devised with two translation steps. The first step targets at the modality-common representations, and leverage adversarial learning to align the inferior non-lexical modalities to the language one. The second one considers the heterogeneous nature of modality-specific representations. Modality-specific representations provides diversity, but they generally convey distinct characteristics that lead to inevitable gap in the feature space. Combining diverse information from different distributions easily pose burden on the fusion network to generalize well on various distributions. To this end, additional distribution constraint is imposed to transfer the 'style' of individual modality-specific representations to the modality-common one. During the two-step translation process, the non-lexical modalities are translated to incorporate with the dominant language modality, and a more uniform feature distribution is obtained both across modalities and across information properties. Note that we only aim to learn a uniform distribution with the same 'style', and meanwhile the diverse contents are preserved to avoid information loss. A more unified cross-modal distribution facilitates multimodal fusion and can cooperate better with simple fusion methods.

As for the decoding and reconstruction stage, we observe the problem of noisy information in previous encoder–decoder-based disentanglement works. To avoid information loss, decoding and feature reconstruction is adopted in many existing works, which helps retain the information contained in the original representation [11–13]. However, strict constraint (e.g., MSE constraint loss) is generally utilized, which easily leads to the preservation of redundant information when the original one is not effective enough. To this end, we propose Slack Reconstruction to apply a more relaxed constraint to the reconstruction process, instead of decoding out reconstructed representation that is exactly the same as the original one. Specifically, we form triplets from the original and reconstructed representations that come from different samples. By regularizing the relationships of different triplets, Slack Reconstruction encourages the reconstructed representation to be more similar to the original one than those from other samples of the same modality. In this way, combining the reconstruction loss and main task loss, we encourage the preservation of task-related sample-specific information, and the decoded representation is still distinguishable with respect to samples. Besides, the more relaxed constraint allows for the difference between the decoded and original representation, which helps reduce the redundancy. Moreover, we explore the inter-sample relationships by learning from various triplets, which can be regarded as a way of data augmentation.

In short, the contributions of this paper can be summarized as follows:

- We propose a novel Disentanglement Translation Network (DTN) for multimodal sentiment analysis. The proposed method disentangles the cross-modal input into desirable properties and learns a unified distribution. Different to previous cross-modal disentanglement learning methods, we devise a two-step translation after unimodal disentanglement, and improve the decoding process to ensure feature quality.
- To narrow down the distributional gap, we propose a two-step translation method, which is incorporated in the disentanglement learning framework and separately considers the modality-common and modality-specific information. The two-step translation achieves finer matching of distribution both across modalities and across information properties.
- To seek a balance between retaining effective information and reducing redundancy, we propose Slack Reconstruction in the decoding and reconstruction stage. Different to the commonly-used strict reconstruction methods that easily preserve redundancy, Slack Reconstruction imposes a more moderate constraint, so that sample-specific information is encouraged to be retained while redundant information can be to some extent reduced in the representations.
- Extensive experiments on two widely-used datasets suggest that DTN achieves competitive performance on the task of multimodal sentiment analysis, which verify the effectiveness of the proposed method.

2. Related work

2.1. Multimodal sentiment analysis

The wide application of smart devices enables the availability of abundant cross-modal data. In recent years, multimodal sentiment

analysis draws increasing attention for its ability to interpret human language and understand latent opinions [6,8,9,14–18], which is considered to be important for intelligent human–computer interaction. Existing works for human multimodal sentiment analysis are mostly dependent on the language, visual and acoustic modalities. Different modalities convey different aspects of individual objects, and the cross-modal information is complementary and interacted. To leverage cross-modal data for more accurately interpreting the latent sentiment, many previous works design different fusion strategies to explore the interactions between modalities [10,16,19–21]. **Early fusion** (a.k.a, feature-level fusion) [17,20,22,23] and **late fusion** (a.k.a, decision-level fusion) [24–27] are two simple and explicit ways applied in early MSA works to fuse different modalities. By combining the cross-modal information, model improvement can be observed compared to the counterparts of only leveraging one single modality.

To fully exploit the multimodal data, a large amount of advanced fusion methods have been proposed to explore the cross-modal dynamics and learn more informative multimodal representation. **Tensor-based methods** [18,28,29] achieve significant improvement for the capacity to model complex cross-modal interaction. For example, [18] applies outer product to learn high-dimensional multimodal tensor, so as to model the unimodal, bimodal and trimodal interactions. **Graph-based fusion methods** [11,30] are also studied, which explore the cross-modal associations across time series and modalities. Moreover, **cross-modal attention mechanisms** [3,10,31,32] are popular methods to explore inter-modal dynamics, which can help identify and highlight important information. [10] takes into account the importance differences between multiple modalities, and assigns weights to them through the importance attention network. To provide interpretability for multimodal fusion, **capsule-based fusion methods** [4,33] are proposed to help understand the connections and associations between different modalities. Besides, to incorporate additional knowledge for better understanding, KnowleNet [34] proposes knowledge fusion network, which leverages ConceptNet knowledge base to utilize prior knowledge and determine image–text relatedness through sample-level and word-level cross-modal semantic similarity detection. Contrastive learning is also introduced to improve the spatial distribution of positive and negative samples. More recently, with the success of Bidirectional Encoder Representations from Transformers (BERT) [35], researchers have explored various ways to fine-tune BERT with cross-modal information [7,36], which shows significant performance improvement on multimodal tasks. For example, MAG-BERT [7] proposes an attachment to allow BERT and XL-Net [37] accept multimodal nonverbal data during fine-tuning, AOBERT [38] proposes a single-stream transformer that can handle three modalities as inputs to one network, and CM-BERT [36] relies on the interaction of text and acoustic modality to fine-tune the pretrained BERT model.

However, the heterogeneous nature of different modalities poses great burden on the fusion process and the learning of multimodal joint embedding. Different modalities are collected from various sensors and convey different aspects of the same object, leading to various characteristics and information properties, which result in distributional discrepancy in the feature space. Besides, it is common that the less-dominant modalities contain redundant information and even much noise, which results in variations and further aggravates the problem of modality gap. Especially when the model adopts simple fusion methods (concatenation, element-wise addition, etc.), the inconsistent distribution of individual modalities brings burden to fuse cross-modal information [12,39], and leads to the difficulty in mining complementary information across modalities [11].

2.2. Disentanglement learning

To bridge the modality gap and learn effective joint representation from various modalities, some methods translate the source modality into the target one to achieve distribution matching. ARGF [11] leverages adversarial learning to translate three modalities to a target distribution, so that modality-invariant distribution can be obtained to narrow down the distributional discrepancy. MTSA [40] translates the other modalities into the text to improve feature quality.

But these methods neglect the unique characteristic of individual modalities from different perspectives [39]. Therefore, some recent works apply disentanglement-based framework to explicitly model the modality-common and modality-specific representations before fusion. Disentanglement learning [12,13,39,41,42] aims to extract desirable information properties from the input data, facilitating better understanding, analysis, and manipulation of the underlying latent factors. It can help bridge the modality gap by extracting shared and independent latent factors that are common across different modalities. By explicitly modeling and separating the desirable information properties, disentanglement learning enables the alignment of cross-modal semantically related concepts, which reduces the modality gap [12,39]. Besides, disentanglement learning also aids multimodal fusion by providing a more structured and explicit representation of the underlying latent factors. By fusing specific desirable information across modalities, multimodal fusion becomes more targeted and effective, leading to improved performance. DisVAE [43] learns disentangled and informative user representations, which accurately characterize user intentions and simultaneously maintain substantial multi-view information. SIMR [13] learns speaker-independent multimodal representation. This framework separates the nonverbal inputs into style encoding and content representation with the aid of informative cross-modal correlations. MISA [12] learns factorized subspaces for each modality and provides better representations as input to fusion. It learns modality-invariant and modality-specific subspaces before fusion, and incorporates a combination of losses to aid learning. FDMER [39] also explores the commonality among different modalities and captures the unique characteristics of each modality. Although different modalities share commonality of the high-level semantic information about the underlying sentiment, each modality may contribute unique subtleties. For example, the language modality can indicate the sentiment tendency by the choice of words and the structure of the sentences, which is not contained in other modalities. Both commonality and diversity across modalities are essential for understanding underlying sentiments, as they provide complementary and interacted information for a comprehensive analysis.

Following previous disentanglement works, we captures the desirable properties (i.e., modality-common and modality-specific information) of the multimodal input to aid fusion and multimodal learning. However, the modality gap problem still exists between information properties, and we propose two-step translation, which is cooperated with the disentanglement learning framework to further unify the distribution both across modalities and across information properties. Besides, to avoid information loss [11] or avoid encoding trivial representations [12], decoding and feature reconstruction is applied in our method. But different to the strict reconstruction as done in previous works [11–13], we explore a more moderate but effective reconstruction method to constrain model learning.

3. Algorithm

In this section, we introduce the proposed method in detail. The diagram of DTN is illustrated in Fig. 2. We focus on learning desirable properties from the cross-modal input and obtaining a uniform distribution, so as to benefit multimodal fusion by alleviating the modality gap problem and reducing redundancy. In this way, the generated representations can cooperate well with even the simple and direct fusion methods. Specifically, DTN conducts disentanglement encoding with a two-step translation, which captures the information shared across modalities and the unique characteristics of individual modalities. In addition, taking into account the problem of information redundancy, we devise slack reconstruction to further benefit learning an effective

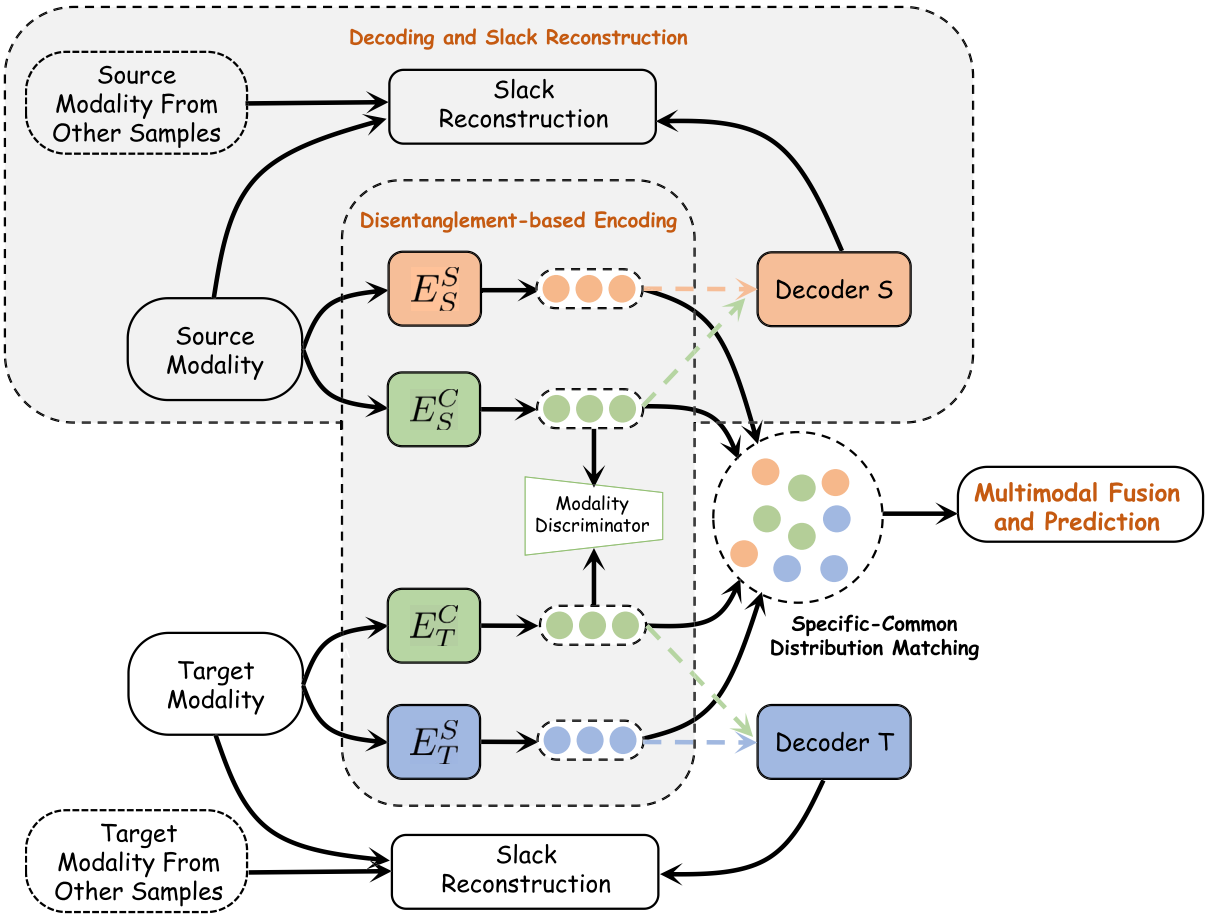

Fig. 2. The diagram of the proposed method. Both the encoding and decoding processes are improved to obtain effective joint multimodal embedding.

joint embedding. As can be seen in Fig. 2, the proposed method is designed based on an encoding–decoding framework with three main parts: disentanglement-based two-step translation, decoding with slack reconstruction, and multimodal fusion. The details of our work are introduced in the following sections.

3.1. Downstream task and unimodal coding

The downstream task in this paper is multimodal sentiment analysis (MSA) with language, visual and audio input. The goal is to score the sentiment intensity to determine the positive/negative sentiment. Before the representation disentanglement, transformer-based models [44] are applied to extract the high-level semantic representations for individual modalities. BERT [35] is utilized as the unimodal coding network for the language modality, while the standard transformer is adopted for the other modalities, which can be formulated as follows:

$$X_m^0 = F^m(U_m; \theta_m), \quad m \in \{l, a, v\} \tag{1}$$

where F^m parameterized by θ^m is the unimodal learning network of modality m, and U_m is the input utterance. To ensure the same feature dimensionality, the temporal convolution operation is applied, which can be formulated as:

$$X_m = \text{Conv 1D}\left(X_m^0, K_m\right) \in \mathbb{R}^{T_m \times d} \tag{2}$$

where Conv 1D denotes the temporal convolution operation and K_m is the kernel size. In this way, we can obtain unimodal representations $X_m \in \mathbb{R}^{T_m \times d}$ $(m \in \{l, a, v\})$ that share the same feature dimensionality, which is more suitable for multimodal fusion.

3.2. Disentanglement-based two-step translation

The heterogeneous nature of individual modalities leads to distributional discrepancy in the feature space, which poses burden on the fusion process [11,12,39]. Besides, the noisy information brings further variations, especially when some of the modalities are much less effective. To address this problem, during the encoding stage, we propose disentanglement-based two-step translation to capture desirable properties of the multimodal input, which alleviates the distributional gap and aids multimodal fusion to obtain informative multimodal representation.

We first decouple the unimodal representations into modality-common and modality-specific subspaces, so as to fully explore the commonality and diversity across modalities, which can be formulated as:

$$X_m^c = E_m^c(X_m) \tag{3}$$

$$X_m^s = E_m^s(X_m) \tag{4}$$

where X_m is the representation of modality m, E_m^c and E_m^s are the modality-common and modality-specific encoders to extract modality-common representation X_m^c and modality-specific representation X_m^s. By doing this, X_m is decoupled into X_m^c and X_m^s. X_m^c conveys the commonality across modalities, while X_m^s reveals the diversity and complementary information. To ensure that the disentangled representation are independent with each other, the irrelevant constraint is applied to orthogonalize the decoupled features. In this way, the feature encoders focus on the desirable properties of each modality, which alleviates the modality gap problem and aids finer multimodal fusion to obtain an effective joint embedding [12,39]. Besides, by leading the encoders to

capture the modality-common and modality-specific information, the encoders tend to be more aware of the task-related information so that noise and redundancy is reduced, which implicitly improves the quality of unimodal and multimodal representations.

However, for multimodal input that involves multiple modalities, the disentangled representations show distributional discrepancy. Take two modalities for example, we first define a source modality s and a target modality t. The gap exists both between modalities (X_t^c and X_s^c) and between information properties (X_s^c and X_s^s). This discrepancy affects the fusion process and the learning of multimodal representation [11]. To this end, we devise a two-step translation and incorporate it with the disentanglement learning framework. The first translation step targets at modality-common representations. We aim to achieve a modality-invariant distribution of X_s^c of the source modality s and that of the target one X_t^c, which is achieved by the min–max game between the modality-common encoders and the modality discriminator:

$$\begin{aligned}\mathcal{L}_c(E_s^c, E_t^c, D) = &\mathbb{E}_{t\sim p(t)}[\log D(E_t^c(t))] \\ &+ \mathbb{E}_{s\sim p(s)}[\log(1 - D(E_s^c(s)))]\end{aligned} \tag{5}$$

where t and s are the input original feature of the target and source modality. E_s^c and E_t^c are the modality-common encoders, and D is the modality discriminator to determine the modality category of the input. By the min–max game between E_s^c, E_t^c and D, the modality-common encoder manage to fool the discriminator, which means that the encoder learns to capture the shared modality-irrelevant information, and X_s^c and X_t^c show similar feature distribution.

As for the second translation step, Specific-Common Distribution Matching is applied, which aims at the modality-specific information. The modality-specific representations provide diversity, which however show distributional gap both between modalities and between information properties. To this end, Specific-Common Constraint is applied to further narrow down the discrepancy. Inspired by the success of adversarial cyclic translation [45] in computer vision tasks, we achieve the distribution transformation of modality-specific representations while at the same time preserving its semantic information:

$$T_{s-c} = T_{s\to c}(X_t^s, X_t^c) + T_{s\to c}(X_s^s, X_t^c) \tag{6}$$

where $T(\cdot)$ denotes the specific-common transformation process. Modality-specific representations from both source and target modalities are led to resemble the distribution of the modality-common one. For the modality-specific representation from each modality, the realization of $T(\cdot)$ can be formulated as:

$$\begin{aligned}\mathcal{L}_{s-c} = &\mathbb{E}_{s\sim p(s)}[\|G_{C\to S}(G_{S\to C}(s)) - s\|_1] \\ &+ \mathbb{E}_{c\sim p(c)}[\|G_{S\to C}(G_{C\to S}(c)) - c\|_1] \\ &+ \mathbb{E}_{c\sim p(c)}[\log D_C(c)] + \mathbb{E}_{s\sim p(s)}[\log(1 - D_C(G_{S\to C}(s)))] \\ &+ \mathbb{E}_{c\sim p(c)}[\log D_S(c)] + \mathbb{E}_{s\sim p(s)}[\log(1 - D_S(G_{C\to S}(s)))]\end{aligned} \tag{7}$$

where S and C are the sets from the modality-specific and modality-common representations, $G_{S\to C}$ is the specific-common generator to transform the distribution of the modality-specific representation to that of the modality-common representation, and $G_{C\to S}$ realizes the inverse transformation from common to specific. D_c and D_s are the discriminators to determine whether the input is common or specific information. In this way $G_{S\to C}$ learns to transform the distribution of modality-specific information to the modality-common one without modifying its semantic contents, and the transformed $\hat{X}_m^s$ can be obtained:

$$\hat{X}_m^s = G_{S\to C}(X_m^s) \tag{8}$$

In this way, for each modality m, both of its modality-common representation X_m^c and its modality-specific representation $\hat{X}_m^s$ are mapped into a similar distribution.

To summarize, Eq. (5) illustrates the first step to learn modality-common representations across modalities, while Eqs. (6)–(8) explain the second step to achieve distribution matching between information properties. By the application of the proposed two-step translation, the distributional gap can be narrowed down considering different aspects in the disentanglement-based framework, and a more uniform feature distribution can be obtained both across modalities and across information properties. Note that during the two-step translation, although the representations are led to have the same distribution, their semantic information is retained.

3.3. Decoding and slack reconstruction

In many previous disentanglement-based learning methods, to avoid information loss in the disentanglement and translation stage, the encoder–decoder structure is adopted. The decoders reconstruct the original representation with the reconstruction constraint. By doing this, information can be effectively preserved when aligning the cross-modal distributions. The decoding operation can be formulated as:

$$X_m^R = D_m(X_m^c, X_m^s) \tag{9}$$

where D_m is the unimodal decoder for modality m, $X_m^R \in \mathbb{R}^{1\times d}$ is the reconstructed representation. The reconstruction loss can be calculated by comparing the difference between the decoded representation X_m^R and the original representation X_m, and the reconstruction loss is applied to constrain model learning so as to lower the risk of information loss. One common way is to calculate the Mean Square Error (MSE) loss between X_m^R and X_m:

$$\mathcal{L}_{recon} = \|X_m - X_m^R\|_2^2 \tag{10}$$

where $\mathcal{L}_{recon}$ is the reconstruction loss. Observing the above equation, it can be seen that MSE is a strict constraint that leads the decoded representation to perfectly resemble the original one. However, the original input may contain redundant or even noisy information that is irrelevant to the downstream task. Especially facing the imbalance of modality dominance, the less-dominant modalities may contain much redundancy or noise, which is useless and may be harmful to the fusion process. And worse, the noisy information may be amplified and may even overwrite the useful one during multimodal fusion, damaging the quality of the multimodal representation. Ideally, during the disentanglement process, the encoders capture the task-related modality-common and modality-specific information, and the desirable information is fused to obtain the multimodal representation for inference. It is likely that the decoded representation is to some extent different to the original one.

To this end, we introduce a more relaxed constraint named Slack Reconstruction to the decoding and reconstruction stage. The idea of Slack Reconstruction can be referred to Fig. 3. Instead of forcing the reconstructed representation to be exactly the same as the original one [11,12], we encourage the reconstructed representation to be more similar to the original one than those from other samples of the same modality. In this way, sample-specific information is retained during the whole encoding and decoding process, so that the decoded representation is still distinguishable with respect to samples. Meanwhile, difference between the decoded and original representations is allowed, which implicitly encourages the reduction of redundant information.

Formally, inspired by previous metric learning methods [46,47], we formulate learning triplets from the original and reconstructed representations. Different to the widely-used triplet loss in computer vision classification tasks, we design the loss function to achieve feature reconstruction in multimodal learning. For each modality, we first reconstruct the original input via the decoding network based on the multimodal representation:

$$X_m^R = D_m(X_M) \tag{11}$$

where X_M is the multimodal representation after fusion and $X_m^R \in \mathbb{R}^{1\times d}$ is the decoded representation of modality m. Different to the input of

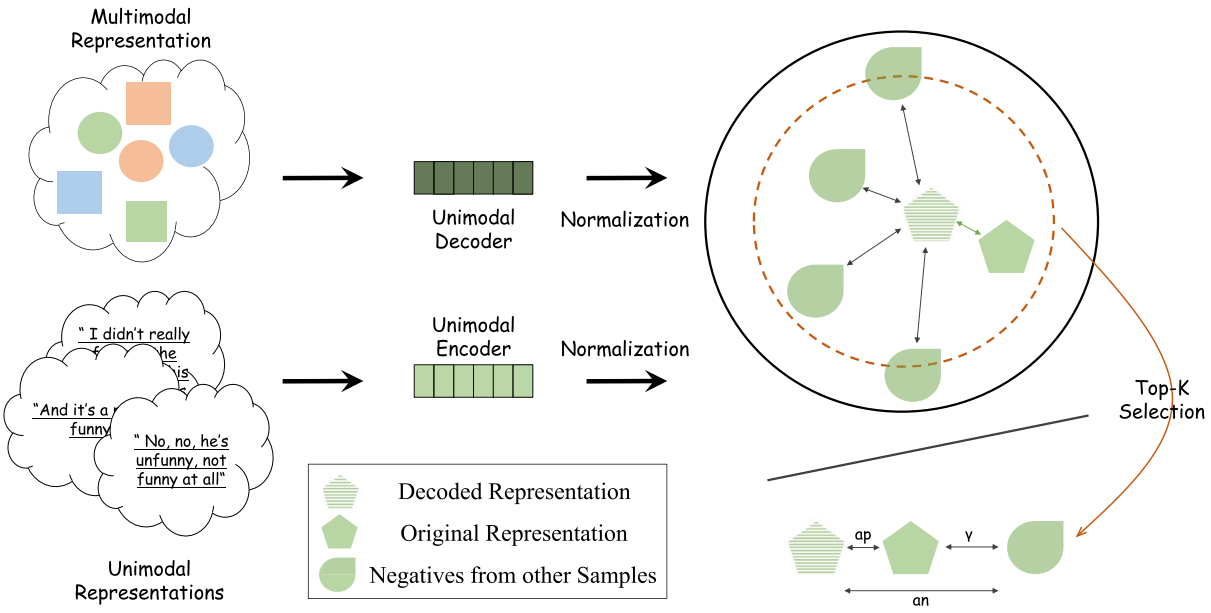

Fig. 3. Slack Reconstruction imposes a more relaxed constraint on the decoding and reconstruction process, so as to retain useful task-related information and meanwhile reduce redundancy.

Eq. (9), we directly decode out the unimodal representations from the multimodal embedding. In this way, we ensure that the multimodal representation contains sufficient desirable knowledge from all modalities. Then, we construct the positive and negative samples for the reconstructed representation (anchor). The positive is obviously the original representation output by the corresponding unimodal coding network as shown in Eq. (1). And we define the hardest samples (the top-k similar representations from other samples in a batch) as the negatives. Afterwards, we define a margin γ which determines the degree of difference between the similarities of anchor-positive and anchor-negative pairs. If the similarity of anchor-positive pair exceeds the similarity of anchor-negative pair by a margin of γ, we deem it as acceptable and the slack reconstruction loss will be equal to zero. For a input batch of n samples $X = [X_m, X_m^N] = [X_m, X_m^1, X_m^2, \ldots, X_m^{n-1}]$, after performing L2-normalization on the representations, the detailed procedures of slack reconstruction loss can be formulated as:

$$S_{AP} = X_m^R (X_m)^T \tag{12}$$

$$S_{AN} = \frac{1}{K} \sum max^K (X_m^R (X_m^N)^T) \tag{13}$$

$$\mathcal{L}_{Recon} = max(S_{AN} - S_{AP} + \gamma, 0) \tag{14}$$

where $X_m^N \in \mathbb{R}^{(n-1)\times d}$ denotes a set of negative samples in the batch, Eq. (12) defines inner product as the similarity between the anchor-positive samples. max^K denotes that we select the top-K anchor-negative pairs with the highest similarity to represent the negative set, which means that Eq. (13) identifies the hardest samples as the negatives for the anchor sample. This operation benefits model learning by paying more attention to the difficult negatives, enhancing the ability to capture sample-wise distinguishable information. And Eq. (14) is the slack reconstruction loss for model update, which considers the anchor-positive–negative relationships from multiple samples.

The proposed Slack Reconstruction seeks a balance between retaining useful information and reducing redundancy. Compared to strict constraint (e.g., MAE or MSE), the decoded representation is led to be more similar to its corresponding original representation but not necessarily to be exactly the same. In this way, combining the function of the reconstruction loss and main task loss, we encourage the preservation of the task-related sample-specific information, and meanwhile allows the reduction of redundancy. Besides, by leveraging multiple negatives, Slack Reconstruction also explores inter-sample relationships during the decoding and reconstruction process. Note that although we form triplets for training, the proposed Slack Reconstruction is different to the triplet loss [47] from various aspects, including the applied task, the construction of learning triplets, and the specific design of the loss function. Unlike traditional triplet loss that typically constructs one positive and one negative for the anchor sample, we leverage multiple negatives to better explore the inter-sample relationships. Moreover, the definition of negative pairs in the design of loss function considers harder samples to enable higher capacity.

3.4. Extension to multiple modalities and model update

Three modalities are involved when the proposed method is applied to the multimodal sentiment analysis task, including language, audio, visual modalities. We denote the input modalities as $m \in \{l, a, v\}$, and the original representations sent to the input are denoted as X_m. Considering that in field of MSA, the language modality plays an dominant role and is the most informative [11,48], we define the language modality as the target modality, while the other two as the source modalities.

For each modality m, its corresponding modality-common/-specific encoders are applied to decouple X_m into X_m^c and X_m^s, respectively, which can be formulated as:

$$X_m^c = E_m^c(X_m) \tag{15}$$

$$X_m^s = E_m^s(X_m) \tag{16}$$

To achieve disentanglement-based translation, we adopt a two-step translation. The first one targets at the cross-modal modality-common representations, aiming to regularize the learning of E_m^c to achieve a uniform distribution for modality-common information X_m^c. As for the second translation step, it aims to achieve distributional matching between modality-common/-specific representations.

For the modality-specific representation X_m^s of each modality, its corresponding specific-common generator $G_{S\to C}^m$ matches its distribution to the modality-common one:

$$\hat{X}_m^s = G_{S\to C}^m(X_m^s) \tag{17}$$

In this way, for the multimodal input we can obtain X_l^c, X_a^c, X_v^c, $\hat{X}_l^s$, $\hat{X}_a^s$ and $\hat{X}_v^s$ with a uniform distribution, which are sent to the fusion module to obtain the multimodal representation. The proposed method

can be integrated with various fusion methods. In our experiments, we adopt the combination of element-wise addition and element-wise multiplication to obtain the joint embedding X_M:

$$X_M = (X_l^c + X_a^c + X_v^c) * (\hat{X}_l^s + \hat{X}_a^s + \hat{X}_v^s) \tag{18}$$

To avoid information loss, decoders are adopted for individual modalities to reconstruct the unimodal representations, and by comparing the difference between the original and decoded representations, reconstruction loss is calculated to impose constraint on model learning. For each modality, we reconstruct its corresponding representation by the unimodal decoder:

$$X_m^R = D_m(X_M) \tag{19}$$

where D_m is the decoder for modality m. Note that we obtain the decoded representations from the multimodal representation, so as to ensure the multimodal embedding contains sufficient useful information from all modalities.

For prediction, the multimodal representation X_M is sent into the multimodal classifier C to draw inference:

$$y_M = C(X^M; \theta_c) \tag{20}$$

where C is the classifier that outputs the prediction y_M.

To achieve better model optimization, we adopt a three-step update strategy to realize targeted optimization for individual components. We first optimize the disentanglement network to ensure decoupling desirable information properties from the cross-modal input, which leverages the irrelevant loss and the loss in the first translation step:

$$\mathcal{L}_1 = \mathcal{L}_{Irre} + \mathcal{L}_C \tag{21}$$

Specifically, to learn modality-specific and modality-common representations that are independent of each other, we orthogonalize the two features and calculate the irrelevant loss $\mathcal{L}_{irre}$, which can be formulated as:

$$\mathcal{L}_{irre} = \langle X_m^c, X_m^s \rangle \tag{22}$$

where $\langle \cdot \rangle$ denotes the inner product operation. When $\mathcal{L}_{irre}$ is being closer to 0, the irrelevance of the domain-specific/-common representations is strengthened.

As for $\mathcal{L}_C$, it constraints the learning of the modality-common encoders to generate a uniform modality-common distribution for cross-modal input. The details can be referred to Eq. (5). For two source modalities a and v, L_C can be formulated as:

$$\mathcal{L}_C = \mathcal{L}_C^a(E_a^c, E_l^c, D) + \mathcal{L}_C^v(E_v^c, E_l^c, D) \tag{23}$$

$$\begin{aligned}\mathcal{L}_C^a(E_a^c, E_l^c, D) = {} & \mathbb{E}_{l\sim p(l)}[\log D(E_l^c(l))] \\ & + \mathbb{E}_{a\sim p(a)}[\log(1 - D(E_a^c(a)))]\end{aligned} \tag{24}$$

$$\begin{aligned}\mathcal{L}_C^v(E_v^c, E_l^c, D) = {} & \mathbb{E}_{l\sim p(l)}[\log D(E_l^c(l))] \\ & + \mathbb{E}_{v\sim p(v)}[\log(1 - D(E_v^c(v)))]\end{aligned} \tag{25}$$

where D is the modality discriminator. In this way, a uniform modality-common distribution can be obtained, which conveys the commonality across modalities.

The second update step is to optimize the generators and discriminators for specific-common matching. During this process, we aim to achieve a specific-common generator for each modality, with which the semantic content of individual modality-specific representations should be preserved, and meanwhile we transform the distributional style of modality-specific representation to the modality-common one. For each modality m, we realize this goal by a cyclic transformation learning way:

$$\begin{aligned}\mathcal{L}_2 = {} & \mathbb{E}_{s\sim p(s)}[\|G_{C\to S}^m(G_{S\to C}^m(s)) - s\|_1] \\ & + \mathbb{E}_{c\sim p(c)}[\|G_{S\to C}^m(G_{C\to S}^m(c)) - c\|_1] \\ & + \mathbb{E}_{c\sim p(c)}[\log D_C(c)] + \mathbb{E}_{s\sim p(s)}[\log(1 - D_C(G_{S\to C}^m(s)))] \\ & + \mathbb{E}_{c\sim p(c)}[\log D_S^m(c)] + \mathbb{E}_{s\sim p(s)}[\log(1 - D_S^m(G_{C\to S}^m(s)))]\end{aligned} \tag{26}$$

where C is the target modality-common distribution and $S \in \{l, a, v\}$ refers to the source modality-specific distribution from all individual modalities. During this process, $\mathcal{L}_2$ is used to update $G_{S\to C}^m(\cdot)$, $G_{C\to S}^m(\cdot)$, D_C and D_S^m. For each modality m, the ultimate goal is to train a specific-common generator $G_{S\to C}^m(\cdot)$ to match the distribution of X_l^s, X_a^s and X_v^s to X_l^c. In this way, more uniform distribution can be obtained without modifying the semantic content, which is beneficial for easier fusion.

The last step is to leverage the prediction and reconstruction loss to update the whole network except for discriminators:

$$\mathcal{L}_3 = \mathcal{L}_{Recon} + \mathcal{L}_{Task} \tag{27}$$

where $\mathcal{L}_{Task}$ is the main task loss, and the details of $\mathcal{L}_{Recon}$ can be referred to Eq. (14). In this way, the main task loss guides the model to focus on task-related information, while the reconstruction loss encourages the preservation of sample-specific information and simultaneously allows for the reduction of redundancy.

To summarize, the proposed DTN explicitly models the desirable properties of the multimodal input, and learns a unified distribution to alleviate the modality gap problem and aid multimodal fusion. Meanwhile, the proposed slack reconstruction further ensures feature effectiveness and improve the decoding and reconstruction process. We also achieve targeted optimization for individual components, so as to maximize the effectiveness of each module and to be more in line with the downstream task.

4. Experiment

4.1. Datasets and evaluation protocols

In this paper, two of the most commonly used public datasets are adopted to carry out experiments, including the CMU-MOSI [27] dataset and CMU-MOSEI [49] dataset for the task of MSA.

(1) **CMU-MOSI** is a widely-used dataset for multimodal sentiment analysis, which is a collection of 2199 opinion video clips. Sentiment intensity is annotated from strongly negative to strongly positive with a linear scale from −3 to +3. To be consistent with prior works, we use 1284 utterances for model training, 229 utterances for validation, and 686 utterances for testing.

(2) **CMU-MOSEI** is a large dataset that consists of 23 454 video utterances from more than 1000 YouTube speakers, covering 250 distinct topics. All the sentences utterance are randomly chosen from various topics and monologue videos, and the sentiment annotation of each also range from −3 to +3. In our experiments, we use 16,265 utterances for model training, 1869 utterances for validation, and 4643 utterances for testing.

In our experiments, five evaluation metrics are adopted to evaluate the model performance, including (1) Acc7: 7-way accuracy, the sentiment score classification; (2) Acc2: binary accuracy, positive or negative classification; (3) F1 score; (4) MAE: mean absolute error and (5) Corr: the correlation of the prediction result. Note that following [7], when calculating the Acc2, F1 score, corr, and MAE, we do not use the neutral utterances.

4.2. Implementation details

The input feature dimensionality of the three modalities are 768, 74, and 47 for the CMU-MOSI dataset. As for CMU-MOSEI dataset, the input dimensionality of text, acoustic, and visual modality are 768, 74, and 35, respectively. We develop our model on the Pytorch framework with CUDA 10.1 and torch 1.4.0. The proposed model is trained with the Adam [50] optimizer with initial learning rate set to 2e−5. For the main task loss, we apply MAE and MSE for the CMU-MOSI and CMU-MOSEI dataset, respectively.

Table 1
Comparison results on CMU-MOSI. The results of the baselines marked with † are obtained from their original papers, and the results of other baselines are reproduced with the same unimodal coding networks.

	Acc7	Acc2	F1	MAE	Corr
TFN-BERT [18]	44.7	82.6	82.6	0.761	0.789
LMF-BERT [29]	45.1	84.0	84.0	0.742	0.785
MULT-BERT [52]	41.5	83.7	83.7	0.767	0.799
GFN-BERT [11]	47.0	84.3	84.3	0.736	0.790
ICCN-BERT [53]	39.0	83.0	83.0	0.860	0.710
MAG-BERT [7]	42.9	83.5	83.5	0.790	0.769
TFR-Net [54]	42.6	84.0	83.9	0.787	0.788
MISA† [12]	42.3	83.4	83.6	0.783	0.761
FDMER† [39]	44.1	84.6	84.7	0.724	0.788
HyCon [14]	46.6	85.2	85.1	0.713	0.790
Self-MM† [55]	–	85.9	85.9	0.713	0.798
MMIM† [56]	46.6	86.0	85.9	**0.700**	0.800
C-MIB [51]	**48.2**	85.2	85.2	0.728	0.793
MIM [57]	47.0	85.9	85.9	0.701	0.805
Ours	48.1	**86.2**	**86.2**	0.714	**0.807**

Table 2
Comparison results on CMU-MOSEI. The results of the baselines marked with † are obtained from their original papers, and the results of other baselines are reproduced with the same unimodal coding networks.

	Acc7	Acc2	F1	MAE	Corr
TFN-BERT [18]	51.8	84.5	84.5	0.622	0.781
LMF-BERT [29]	51.2	84.2	84.3	0.612	0.779
MULT-BERT [52]	50.7	84.7	84.6	0.625	0.775
GFN-BERT [11]	51.8	85.0	85.0	0.611	0.774
ICCN-BERT [53]	51.6	84.2	84.2	0.565	0.713
MAG-BERT [7]	51.9	85.0	85.0	0.602	0.778
TFR-Net [54]	51.7	85.2	85.1	0.606	0.781
MISA† [12]	52.2	85.5	85.3	0.555	0.756
FDMER† [39]	54.1	86.1	85.8	0.536	0.773
HyCon [14]	52.8	85.4	85.6	0.601	0.776
Self-MM† [55]	–	85.1	85.3	0.530	0.765
MMIM† [56]	**54.2**	85.9	85.9	**0.526**	0.772
C-MIB [51]	53.0	86.2	86.2	0.584	0.789
MIM [57]	52.5	**86.4**	**86.3**	0.579	**0.792**
Ours	52.5	86.3	**86.3**	0.579	0.788

Table 3
Ablation studies on the CMU-MOSI dataset.

	Acc7	Acc2	F1	MAE	Corr
w/o Disentanglement	45.1	85.1	85.1	0.743	0.790
w/o M_C Translation	45.9	84.8	84.8	0.718	0.793
w/o Irrelevant Constraint	45.1	85.6	85.5	0.727	0.792
w/o S-C Distribution Matching	43.3	85.9	85.9	0.785	0.794
w/o Slack Reconstruction	45.5	85.3	85.2	0.708	0.799
Ours	**48.1**	**86.2**	**86.2**	**0.714**	**0.807**

4.3. Comparison with baseline methods

In this section, we compare the proposed model with other methods on the task of MSA. Experiments are performed on the CMU-MOSI and CMU-MOSEI datasets. As shown in Tables 1 and 2, the proposed method outperforms all baselines on most evaluation metrics on both datasets. The best results are highlighted in bold, and the second best results are marked with underline. Specifically, on the CMU-MOSI dataset, the proposed method achieves 86.2% and 48.1% on Acc2 and Acc7, and outperforms the compared methods on other metrics. While on the CMU-MOSEI dataset, DTN achieves 86.3% Acc2 and 0.788 Corr. The improvement is considerable as human performance is merely around 85% on Acc2 [18]. Compared to recent disentanglement-based methods [12,39], the proposed method shows consistent improvement on both datasets. Specifically, compared to MISA [12] which also realizes disentanglement learning with the encoder–decoder framework, the proposed DTN achieves a significant improvement of 2.8% on Acc2. We argue that the disentanglement learning framework combined with adversarial learning shows higher capacity to model the two desirable information properties and transform distribution. Besides, the decoding and reconstruction process in MISA adopts strict constraint, which inevitably retains the redundant or noisy information (especially for the audio and visual modalities which are much less effective). On the contrary, the Slack Reconstruction in our work seeks a balance between retaining sample-specific effective information and reducing information redundancy. In this way, we implicitly improve the effectiveness of the mulitmodal representation for more accurate prediction. And different to FDMER [39] which also leverages adversarial learning in the disentanglement process, we analyze that our improvement can be credited to the proposed two-step translation, which achieves more targeted distributional transformation for the decoupled representations. And a more unified distribution is obtained both across modalities and across information properties, which better aids multimodal fusion and the subsequent learning. While compared to Multimodal information bottleneck (MIB) [51] that tackles the noisy modality problem with the information bottleneck theory, the proposed method achieves competitive results. MIB explicitly filters out the target-irrelevant information in unimodal representations, which may lead to the loss of modality-specific information. Unlike it, DTN preserves both of the modality-common and modality-specific information during the framework design.

Notably, the proposed method achieves satisfactory results with simple fusion methods (i.e., combining element-wise addition and multiplication). We analyze that the proposed DTN explicitly models the modality-common and modality-specific representations of each modality, and both the shared and complementary information is transformed to a uniform distribution. This operation facilitates the learning of desirable information properties, which aids multimodal fusion even cooperated with simple fusion methods. Moreover, the devised Slack Reconstruction reduces information redundancy in the generated representations, which lowers the role of redundant and noisy information in the fusion process and improves the quality of multimodal representation.

4.4. Ablation study

In this section, to investigate the effectiveness and importance of each component in the proposed DTN, we conduct ablation experiments with experimental results presented in Table 3, and the corresponding analysis are as follows:

(1) Disentanglement Learning: In the case of 'w/o Disentanglement', we remove the whole two-step translation from the model. Performance degradation can be observed on all evaluation metrics. The results indicate the effectiveness of the proposed disentanglement learning framework combined with the two-step translation, without which the model fails to explicitly learn the desirable properties of the cross-modal data. Without the two-step translation, the disentanglement learning in the network becomes a useless step, which still tries to extract useful features but is not effective enough. Applying the two-step translation to the disentanglement network fully enables the disentanglement encoders to capture modality-common and modality-specific information, ensuring the model capacity to sufficiently learn from the complex cross-modal data.

(2) Modality-Common Translation: In the case of 'w/o M_C Translation', we remove the first translation step, which is used to lead the modality-common encoder to capture shared information across multiple modalities. It can be seen that there is a 2.2% drop on Acc7 and 1.4% drop on Acc2. The results are as expected as the first translation step is an essential part to constrain disentanglement learning. Without it, distributional gap exists between cross-modal modality-common representations, and it is uncertain about the output of the encoders, affecting feature quality and the subsequent learning.

Table 4
Discussion on reconstruction methods.

	Acc7	Acc2	F1	MAE	Corr
w/MAE	45.5	85.3	85.2	0.730	0.797
w/MSE	44.5	86.1	85.9	**0.708**	0.790
w/Triplet	45.6	85.6	85.6	0.716	0.805
w/Slack Reconstruction	**48.1**	**86.2**	**86.2**	0.714	**0.807**

Table 5
Discussion on the fusion strategies on CMU-MOSI dataset. Experiments are performed on three simple fusion strategies.

	Acc7	Acc2	F1	MAE	Corr
Addition	47.4	86.1	86.0	0.707	0.799
Concatenation	46.4	85.8	85.7	**0.704**	0.805
Addition+Multiplication	**48.1**	**86.2**	**86.2**	0.714	**0.807**

(3) Irrelevant Constraint: In the case of 'w/o M_C Translation', we remove the irrelevant loss in the disentanglement network, which aims to strengthen the irrelevance between the modality-common and modality-specific information. Based on the results, there is a drop in model performance but not significant, which is reasonable. Without the irrelevant loss, the two representations may be correlated. In this case, the two information properties may not be well projected into two subspaces. But it seems that it is not a severe problem, mainly due to the reason that the useful information is still preserved.

(4) Specific-Common Distribution Matching: In the case of 'w/o S-C Distribution Matching', we aim to verify the second translation step of unifying the distribution of the modality-specific and modality-common representations. When it is removed from the model, there is a slight drop on the evaluation metrics of Acc2, F1 and Corr. As modality-specific representations serves to convey diversity and complementary information, the gap between information properties may not lead to severe degradation. Still, if it is applied, all evaluation metrics show improvement. We believe that S-C Distribution Matching helps unify the distribution across modalities and across information properties, which aids finer fusion and can cooperate well with simple fusion methods.

(5) Slack Reconstruction: In the case of 'w/o Slack Reconstruction', we investigate the contribution of Slack Reconstruction. If Slack Reconstruction is removed from the network, the model performance deteriorates. It may be due to the reason that without the reconstruction constraint, there may be information loss or the encoders may encode trivial information. The reconstruction process enables the preservation of useful information, which implicitly ensure representation quality during the whole encoding stage

4.5. Discussion on reconstruction methods

In this section, we conduct experiments to further verify the effectiveness of the proposed Slack Reconstruction method. We first compare it with other widely-adopted strict reconstruction constraints. The experimental results can be referred to Table 4, where 'MAE' denotes that we calculate the MAE loss between the reconstructed feature and the original feature, 'MSE' means that we replace the reconstruction constraint with the MSE between the reconstructed feature and the original feature. It can be seen that applying the reconstruction loss improves model performance compared to the case without reconstruction loss (see the case of 'w/o Slack Reconstruction' in Table 3), and the proposed Slack Reconstruction method outperforms both the MAE and MSE reconstruction constraints significantly. Moreover, we observe that applying MSE for reconstruction shows better performance than MAE, which achieves a 86.1% Acc2 and 85.9% F1. We analyze that MSE achieves better reconstruction by paying more attention to the outliers, imposing better constraint to the model to retain useful information. The idea of the top-k negative selection of Slack Reconstruction shares similar idea to be more focused on the hard examples. However, applying Slack Reconstruction still shows improvement on most of the evaluation metrics, especially on Acc7 and MAE. We analyze that the strict reconstruction loss imposes strict constraint between the decoded and original feature. Although this operation helps avoid information loss, the redundant information may be retained in the encoding stage and the reconstructed representation. The problem may become more severe when some of the modalities are less effective and contain much redundancy. The redundant or even noisy information still exists and affects multimodal learning. On the contrary, Slack Reconstruction imposes a more relaxed constraint. It decodes out unimodal features that are sample-wise distinguishable in the feature space, but are not necessarily the same as the original one. In this way, the model components pay more attention on the necessary sample-related information.

Also note that the proposed Slack Reconstruction loss is different to the widely-used triplet loss, which similarly forms learning triplets. And we compare the two loss functions as presented in Table 4 (see the case of 'w/Triplet'). For the traditional triplet loss, we construct 'anchor, positive, negative' triplets from the original and reconstructed unimodal representations. For each training sample, we regard the reconstructed representation as the anchor, and the original input as the positive. While for the negative one, we randomly select one of other samples in the same mini-batch. Based on the experimental results, it can be seen that reconstruction with the traditional triplet loss achieves similar performance to that with the MSE loss, but it still performs worse than the proposed Slack Reconstruction on most of the evaluation metrics. Although both of them construct triplets for model training, Slack reconstruction explores the relationships between various samples, which facilitates the learning of sample-specific useful information. Besides, the proper attention on harder samples further helps improve model capacity.

4.6. Analysis on fusion methods

In this section, we aim to verify the generalization ability of the proposed DTN with different fusion strategies. Multiple fusion strategies are considered, including addition, concatenation, multiplication and their combinations, which can be formulated as:

(1) **Addition** performs element-wise addition of the decoupled unimodal representations to obtain the multimodal representation, which can be formulated as:

$$X_M = X_l^c + X_a^c + X_v^c + \hat{X}_l^s + \hat{X}_a^s + \hat{X}_v^s \tag{28}$$

(2) **Concatenation** concatenates the six decoupled unimodal and obtains a multimodal representation of $\mathbb{R}^{6\times d}$. To ensure the multimodal representation has the same feature dimensionality as d, a fully-connected layer is applied after the concatenation operation, which can be formulated as:

$$X_M = X_l^c \oplus X_a^c \oplus X_v^c \oplus \hat{X}_l^s \oplus \hat{X}_a^s \oplus \hat{X}_v^s \tag{29}$$

(3) **Addition+Multiplication** combines element-wise addition with element-wise multiplication of the decoupled unimodal representations, which can be formulated as:

$$X_M = (X_l^c + X_a^c + X_v^c) * (\hat{X}_l^s + \hat{X}_a^s + \hat{X}_v^s) \tag{30}$$

From the experimental results presented in Table 5, it can be seen that the proposed DTN can consistently achieve satisfactory performance when different fusion strategies are applied. As shown in Table 5, even with simple fusion methods, DTN outperforms the compared baseline methods in most cases. A conclusion can be drawn that the design of DTN is effective and is of satisfactory generalization ability to be applied with various fusion strategies. Besides, comparing the performance between different fusion strategies, three simple methods achieve close results, but fusion with combining addition and

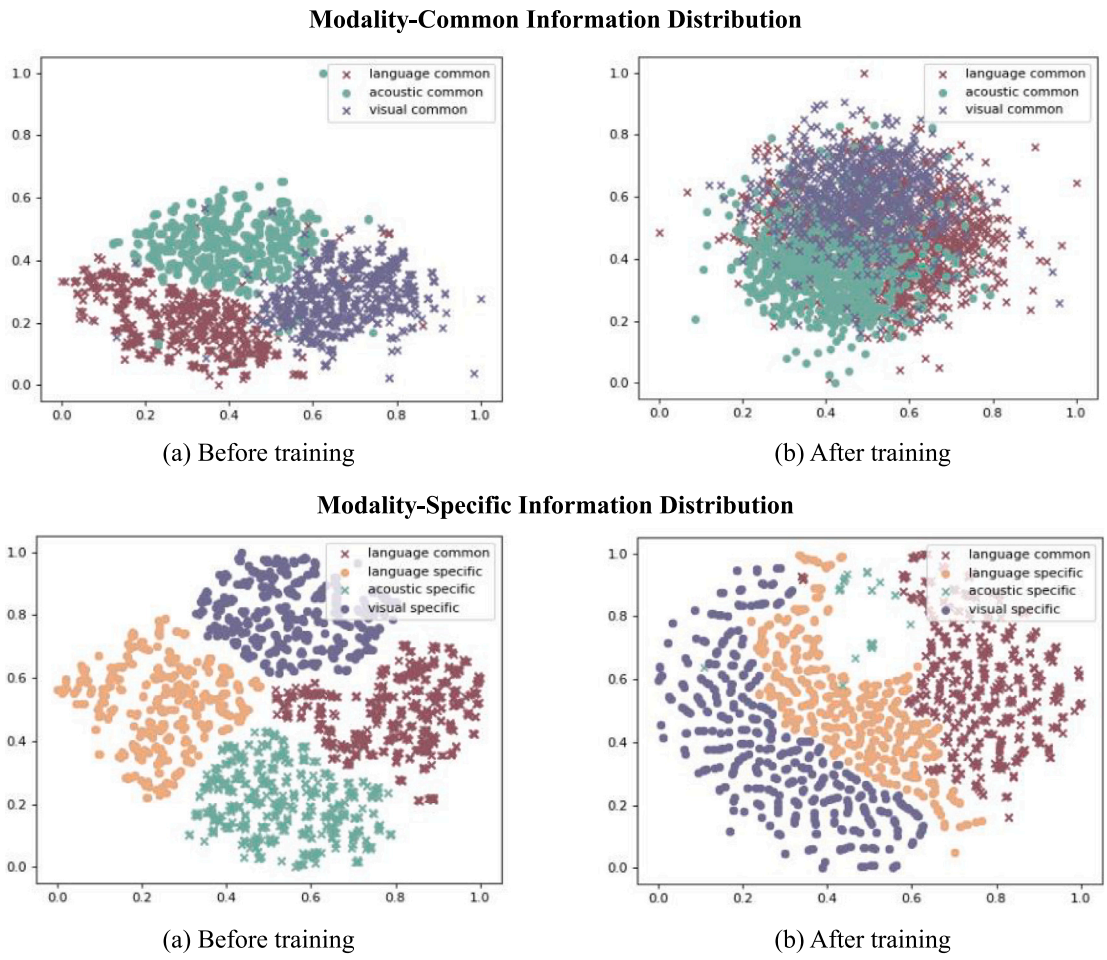

Fig. 4. Visualization of the model effect. (a) The modality-common and modality-specific information before training. (b) The modality-common and modality-specific information after training, which shows a more unified distribution.

multiplication performs the best and is slightly better than others. Still, the proposed method can cooperate with other combinations of fusion methods. And if working with more complex fusion methods, it may gain higher improvement on model performance.

4.7. Visualization of feature space

As shown in Fig. 4, we visualize the modality-common and modality-specific representations before and after training, so as to provide a more intuitive illustration of the effect of the proposed model. The visualization is conducted by using T-SNE [58] with samples from the testing set of the CMU-MOSI dataset.

For the modality-common representations, it can be seen that before training, representations from different modalities are clustered into different clusters with clear boundary in the feature space, which means that features from different modalities show distinctive characteristics and distributional discrepancy. Instead, after training, the modality-common representations from individual modalities are scattered in the feature space, indicating that the encoders capture commonality shared across modalities.

As for the modality-specific representations, there is a clear boundary between each of them and the language common representation before model training. On the contrary, after training, the representations tend to be more scattered in the feature space, and representations from different modalities become closer and show more similar distribution. Note that we also provide the visualization of the language-common representation for better illustration. The proposed method aims to achieve a uniform distribution both across modalities and across information properties. The visualization of both the modality-common and modality-specific information verifies the effectiveness of the proposed method. Besides, we observe that for those modality-specific representations from different modalities and information properties, although distribution matching is performed, the discrepancy cannot be completely eliminated. We analyze that the original modality-specific representations shows more severe distributional gap as they convey the diversity of individual modalities. Still, by the specific-common distribution matching operation, the modality-specific representations shows more similar distribution.

The visualization experiments verify the effectiveness of the two-step translation, and we can see that the proposed method manages to learn a unified distribution for the complex cross-modal input.

5. Conclusion

In this paper, we propose a Disentanglement Translation Network with Slack Reconstruction to improve existing multimodal disentanglement-based learning methods. Disentanglement-based Translation and Slack Reconstruction improve both the encoding and decoding process to achieve more unified distribution with less redundancy. In the encoding stage, two-step translation cooperated with disentanglement learning realizes the learning of desirable information properties and the generation of a more unified distribution, aiding multimodal fusion and obtaining more effective representations. As for the decoding and reconstruction stage, unlike previous strict reconstruction methods, the devised Slack Reconstruction achieves a

more moderate constraint on model learning, and seeks a balance between retaining effective sample-wise task-related information and reducing useless redundancy. Extensive experiments demonstrate the effectiveness of the proposed method.

CRediT authorship contribution statement

Ying Zeng: Conceptualization, Methodology, Experiment, Writing, Revised the manuscript. **Wenjun Yan:** Conceptualization, Methodology. **Sijie Mai:** Conceptualization, Writing. **Haifeng Hu:** Conceptualization, Revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62076262, 61673402, 61273270, 60802069).

References

[1] C. Lee, M. Schaar, A variational information bottleneck approach to multi-omics data integration, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1513–1521.
[2] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, 2017, arXiv preprint arXiv:1705.06950.
[3] A. Shenoy, A. Sardana, N. Graphics, Multilogue-net: A context aware RNN for multi-modal emotion detection and sentiment analysis in conversation, in: ACL 2020, 2020, p. 19.
[4] Y.-H.H. Tsai, M.Q. Ma, M. Yang, R. Salakhutdinov, L.-P. Morency, Multimodal routing: Improving local and global interpretability of multimodal language analysis, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, NIH Public Access, 2020, p. 1823.
[5] Q. Li, D. Gkoumas, C. Lioma, M. Melucci, Quantum-inspired multimodal fusion for video sentiment analysis, Inf. Fusion 65 (2021) 58–71.
[6] Y. Zeng, S. Mai, H. Hu, Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1262–1274.
[7] W. Rahman, M. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for Computational Linguistics. Meeting 2020, 2020, pp. 2359–2369.
[8] F. Zhang, X.-C. Li, C.P. Lim, Q. Hua, C.-R. Dong, J.-H. Zhai, Deep emotional arousal network for multimodal sentiment analysis and emotion recognition, Inf. Fusion 88 (2022) 296–304.
[9] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, X. Kong, Multimodal sentiment analysis based on fusion methods: A survey, Inf. Fusion 95 (2023) 306–325.
[10] S. Liu, P. Gao, Y. Li, W. Fu, W. Ding, Multi-modal fusion network with complementarity and importance for emotion recognition, Inform. Sci. 619 (2023) 679–694.
[11] S. Mai, H. Hu, S. Xing, Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 164–172.
[12] D. Hazarika, R. Zimmermann, S. Poria, Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.
[13] J. Wang, S. Wang, M. Lin, Z. Xu, W. Guo, Learning speaker-independent multimodal representation for sentiment analysis, Inform. Sci. 628 (2023) 208–225.
[14] S. Mai, Y. Zeng, S. Zheng, H. Hu, Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis, IEEE Trans. Affect. Comput. (2022).
[15] Z. Li, Q. Guo, Y. Pan, W. Ding, J. Yu, Y. Zhang, W. Liu, H. Chen, H. Wang, Y. Xie, Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis, Inf. Fusion (2023) 101891.
[16] S. Mai, H. Hu, S. Xing, A unimodal representation learning and recurrent decomposition fusion structure for utterance-level multimodal embedding learning, IEEE Trans. Multimed. 24 (2021) 2488–2501.
[17] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, L.P. Morency, Context-dependent sentiment analysis in user-generated videos, in: ACL, 2017, pp. 873–883.
[18] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.P. Morency, Tensor fusion network for multimodal sentiment analysis, in: EMNLP, 2017, pp. 1114–1125.
[19] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[20] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: Proceedings of IEEE International Conference on Data Mining (ICDM), 2016, pp. 439–448.
[21] L. Pang, S. Zhu, C.W. Ngo, Deep multimodal learning for affective analysis and retrieval, IEEE Trans. Multimed. 17 (11) (2015) 2008–2020.
[22] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.P. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.
[23] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, R. Prasad, Ensemble of SVM trees for multimodal emotion recognition, in: Signal and Information Processing Association Summit and Conference, 2012, pp. 1–4.
[24] C.H. Wu, W.B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Trans. Affect. Comput. 2 (1) (2010) 10–21.
[25] B. Nojavanasghari, D. Gopinath, J. Koushik, L.P. Morency, Deep multimodal fusion for persuasiveness prediction, in: Proceedings of ACM International Conference on Multimodal Interaction, 2016, pp. 284–288.
[26] O. Kampman, D. Bertero, P.N. Fung, et al., Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018): Short Papers, 2018, p. 606.
[27] A. Zadeh, R. Zellers, E. Pincus, L.P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.
[28] P.P. Liang, Z. Liu, Y.-H.H. Tsai, Q. Zhao, R. Salakhutdinov, L.-P. Morency, Learning representations from imperfect time series data via tensor rank regularization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1569–1576.
[29] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A.B. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2247–2256.
[30] Z. Kang, G. Shi, S. Huang, W. Chen, X. Pu, J.T. Zhou, Z. Xu, Multi-graph fusion for multi-view spectral clustering, Knowl.-Based Syst. 189 (2020) 105102.
[31] M.S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, P. Bhattacharyya, Multi-task learning for multi-modal emotion recognition and sentiment analysis, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 370–379.
[32] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016, pp. 289–297.
[33] J. Wu, S. Mai, H. Hu, Graph capsule aggregation for unaligned multimodal sequences, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 521–529.
[34] T. Yue, R. Mao, H. Wang, Z. Hu, E. Cambria, KnowleNet: Knowledge fusion network for multimodal sarcasm detection, Inf. Fusion (2023) 101921.
[35] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[36] K. Yang, H. Xu, K. Gao, CM-BERT: Cross-modal BERT for text-audio sentiment analysis, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 521–528.
[37] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q.V. Le, Xlnet: generalized autoregressive pretraining for language understanding, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 5753–5763.
[38] K. Kim, S. Park, AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis, Inf. Fusion 92 (2023) 37–45.
[39] D. Yang, S. Huang, H. Kuang, Y. Du, L. Zhang, Disentangled representation learning for multimodal emotion recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1642–1651.
[40] B. Yang, B. Shao, L. Wu, X. Lin, Multimodal sentiment analysis with unidirectional modality translation, Neurocomputing 467 (2022) 130–137.
[41] W. Yan, Y. Zeng, H. Hu, Domain adversarial disentanglement network with cross-domain synthesis for generalized face anti-spoofing, IEEE Trans. Circuits Syst. Video Technol. 32 (10) (2022) 7033–7046.

[42] Y. Li, Y. Lu, M. Gong, L. Liu, L. Zhao, Dual-channel feature disentanglement for identity-invariant facial expression recognition, Inform. Sci. 608 (2022) 410–423.
[43] W. Tang, B. Hui, L. Tian, G. Luo, Z. He, Z. Cai, Learning disentangled user representation with multi-view information fusion on social networks, Inf. Fusion 74 (2021) 77–86.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[45] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[46] K.Q. Weinberger, Distance metric learning for large margin nearest neighbor classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[47] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[48] H. Pham, P.P. Liang, T. Manzini, L.P. Morency, P. Barnabăs, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: AAAI, 2019, pp. 6892–6899.
[49] A. Zadeh, P.P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, L.P. Morency, Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: ACL, 2018, pp. 2236–2246.
[50] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[51] S. Mai, Y. Zeng, H. Hu, Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations, IEEE Trans. Multimed. (2022) 1, http://dx.doi.org/10.1109/TMM.2022.3171679.
[52] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting 2019, NIH Public Access, 2019, p. 6558.
[53] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 8992–8999.
[54] Z. Yuan, W. Li, H. Xu, W. Yu, Transformer-based feature reconstruction network for robust multimodal sentiment analysis, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4400–4407.
[55] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 10790–10797.
[56] W. Han, H. Chen, S. Poria, Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 9180–9192.
[57] Y. Zeng, S. Mai, W. Yan, H. Hu, Multimodal reaction: Information modulation for cross-modal representation learning, IEEE Trans. Multimed. (2023) 1–14, http://dx.doi.org/10.1109/TMM.2023.3293335.
[58] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.