Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations

Sijie Mai[1], Ying Zeng[1], Haifeng Hu

Abstract—Learning effective joint embedding for cross-modal data has always been a focus in the field of multimodal machine learning. We argue that during multimodal fusion, the generated multimodal embedding may be redundant, and the discriminative unimodal information may be ignored, which often interferes with accurate prediction and leads to a higher risk of overfitting. Moreover, unimodal representations also contain noisy information that negatively influences the learning of cross-modal dynamics. To this end, we introduce the multimodal information bottleneck (MIB), aiming to learn a powerful and sufficient multimodal representation that is free of redundancy and to filter out noisy information in unimodal representations. Specifically, inheriting from the general information bottleneck (IB), MIB aims to learn the minimal sufficient representation for a given task by maximizing the mutual information between the representation and the target and simultaneously constraining the mutual information between the representation and the input data. Different from general IB, our MIB regularizes both the multimodal and unimodal representations, which is a comprehensive and flexible framework that is compatible with any fusion methods. We develop three MIB variants, namely, early-fusion MIB, late-fusion MIB, and complete MIB, to focus on different perspectives of information constraints. Experimental results suggest that the proposed method reaches state-of-the-art performance on the tasks of multimodal sentiment analysis and multimodal emotion recognition across three widely used datasets. The codes are available at https://github.com/TmacMai/Multimodal-Information-Bottleneck.

Index Terms—multimodal sentiment analysis, information bottleneck, multimodal emotion recognition, representation learning

I. Introduction

With the growing trend of sharing lives on social media, videos that express human opinion are easily accessed, which is worthwhile for the study of understanding the implicit meanings. In those videos, people express their views and emotions through multiple modalities. Apart from spoken language, facial expressions and acoustic clues are also informative during communication between humans or between humans and machines [1]–[3]. The three modalities are often complementary and interact with each other. Consequently, a multimodal learning architecture is necessary for a comprehensive interpretation of human language.

Previous works have tended to design complex fusion strategies to learn multimodal embedding with unimodal features, such as tensor fusion [4] and transformer-based fusion [5], [6]. Though effective, these methods suffer from high complexity, and the generated high-dimensional multimodal embedding is inevitably redundant, which leads to a higher risk of overfitting. Ideally, multimodal embedding should (1) contain the maximum information that is necessary for correct prediction and (2) contain the least information that is irrelevant to the prediction. However, it is not straightforward to learn such perfect embedding. We argue that the multimodal embedding generated by the complex fusion network may be redundant such that the discriminative unimodal information is ignored. For example, Zadeh et al. [4] used an outer product to generate a high-order multimodal tensor, which produces redundant representations that may overwrite the useful unimodal information during prediction.

Moreover, unimodal representations inevitably contain noisy information, which may negatively influence correct prediction. In the field of multimodal sentiment analysis and emotion recognition, previous studies [3], [7] have demonstrated that the language modality is always dominant, while the visual and acoustic modalities are more like auxiliary modalities. These nonlexical modalities often contain noise that may interfere with the learning of cross-modal interactions and influence the quality of the multimodal representation.

How can we ensure that the learned latent representation is sufficient for correct inference and simultaneously free of redundancy and noise? In this paper, drawing inspiration from the information bottleneck [8]–[10], we propose a multimodal information bottleneck (MIB) to learn such unimodal and multimodal representations. The information bottleneck (IB) is based on mutual information (MI), aiming to maximize the MI between the encoded representations and the labels and simultaneously minimize the MI between the encoded representation and the input. IB is able to find the concise and compressed representation of the input, taking into account the complexity of a signal from an information theory perspective [8]. By applying the IB principle, the model can learn to filter out noisy and redundant information

[1]These authors contribute equally.

that might interfere with the prediction and thus obtain minimal sufficient representation for prediction. In our work, IB is used to extract concise representation from cross-modal embedding that discards the irrelevant information to the target. A multimodal sample carries considerable information, and the multimodal embedding generated by complicated fusion methods suffers from high complexity and high dimensionality, which easily leads to redundancy. Therefore, extracting concise representations for multimodal learning is significant.

Specifically, to reduce the redundant information in the generated multimodal representation, we propose an early-fusion MIB (E-MIB). The early fusion here denotes that we first generate the primary multimodal representation via fusion and then, based on it, learn the minimal sufficient multimodal representation via the IB principle (the early fusion here is different from the commonly considered feature concatenation). First, we extract and fuse the high-level unimodal representations of the three modalities to obtain the primary joint multimodal representation, where the fusion method is flexible. Then, we feed the primary joint multimodal representation to the information bottleneck, based on which the multimodal representation is distilled and compressed. Ideally, the compressed joint representation only contains information that is relevant to the task, and noisy or redundant information is filtered out. The compressed multimodal representation is therefore free of redundancy, and the risk of overfitting can be reduced.

Moreover, to filter out noise that is irrelevant to the prediction in the unimodal representation, we propose a late-fusion MIB (L-MIB). The late fusion here denotes that we first leverage the IB principle to learn the minimal sufficient unimodal representations and based on them generate the multimodal representation via fusion (the late fusion here is different from the commonly considered late fusion, i.e. voting based on unimodal predictions). By this means, the extracted concise unimodal representation is free of noise, and the generated multimodal representation can be more discriminative. In addition, by applying the IB principle, the distribution gap between modalities can be reduced, such that they can be better fused.

Last but not least, to comprehensively inject the IB principle into multimodal systems, E-MIB and L-MIB are combined to construct the complete MIB framework (C-MIB). C-MIB takes into account filtering out noise in the unimodal representations as well as learning a minimal sufficient multimodal representation, which is expected to be more comprehensive. We compare the performance of L-MIB, E-MIB and C-MIB in Section IV-F. Note that our MIB framework is highly flexible, which is compatible with any fusion mechanisms to provide higher expressive power. The comparison of fusion mechanisms is also shown in Section IV-F.

In conclusion, we propose a new architecture named the multimodal information bottleneck (MIB) to address the multimodal learning problem. The contributions are listed below:

- We propose a novel architecture to model multimodal signals. We innovate to inject the IB principle into multimodal learning to filter out noisy information and reduce redundancy. To the best of our knowledge, we are the first to leverage the IB principle to learn both minimal sufficient unimodal and multimodal representations in multimodal sentiment analysis and emotion recognition.
- E-MIB and L-MIB are proposed to learn minimal sufficient multimodal representation and filter out the noise in unimodal representations, respectively. Combining the advantages of E-MIB and L-MIB, C-MIB is proposed to comprehensively learn more favorable unimodal and multimodal representations. Our MIB framework is generalized to be applied with commonly used fusion strategies to provide higher expressive power.
- The MIB variants outperform other methods on multiple datasets. In addition, the visualization experiments demonstrate that our MIB enables the multimodal representations to be more discriminative. Moreover, our method can reach state-of-the-art performance even with very simple fusion methods.

II. Related Work

A. Multimodal sentiment analysis and emotion recognition

Multimodal sentiment analysis and emotion recognition have attracted significant research attention recently [2]. Previous publications focus on learning various fusion strategies to explore inter-modal dynamics [11]–[14]. One of the simplest ways is to concatenate features at input feature level, which demonstrates improvement over single modality [13], [15], [16]. In contrast, a large number of prior works firstly infer prediction according to each modality and combine the predictions from all modalities using some voting mechanisms [17]–[20]. Nevertheless, these two types of fusion methods cannot explicitly model cross-modal interactions [4].

Consequently, more advanced fusion strategies have been proposed in the past few years. Specifically, performing tensor fusion has received much attention [21]–[23]. Tensor Fusion Network (TFN) [4] adopts outer product to learn joint representation of three modalities. However, the generated multimodal representation is high-dimensional and of high-complexity, which contains redundant information and may suffer from overfitting. Furthermore, modality-translation methods such as Multimodal Cyclic Translation Network (MCTN) [24] and Multimodal Transformer [6] aim at learning joint representation by translating source modality into the target one. More recently, graph-based fusion methods have emerged, which can effectively extract interaction across time series or modalities [25]–[27]. For example, Graph-MFN [26] uses a dynamic fusion graph to fuse features, and Mai et al. [25] conduct fusion using a graph fusion network which regards each modality as one node. Furthermore,

using cross-modal attention mechanism to explore inter-modal dynamics are also very popular [28], [29]. To avoid sarcasm and ambiguity, many methods align the three modalities in the time dimension and then fuse the unimodal features at word level [1], [3], [30], [31]. For instance, Recurrent Attended Variation Embedding Network (RAVEN) [30] models multimodal language by shifting word representations based on audio and vision clues at word level. Moreover, aiming to address the interpretable problem in multimodal fusion, quantum-inspired fusion [32] and capsule-based fusion [33] are emerged.

More recently, with the success of Bidirectional Encoder Representations from Transformers (BERT) [34], fine-tuning BERT has obtained new state-of-the-art results on many natural language processing (NLP) tasks. The original BERT is a pretrained language model, which only accepts lexical information. MAG-BERT [35] proposes an attachment, namely Multimodal Adaptative Gate (MAG), to allow BERT and XL-Net [36] to accept multimodal nonverbal data during fine-tuning. Cross-Modal BERT (CM-BERT) [37] relies on the interaction of text and audio modality to fine-tune the pretrained BERT model, where a masked multimodal attention is designed to dynamically adjust the weight of words by combining the combination of text and audio modality.

Different from the above works, our MIB focuses on learning the minimal sufficient representation instead of designing sophisticated fusion strategies, which is flexible as it can integrate any fusion mechanisms. The proposed MIB is elegant and effective, which can filter out noisy information and reduce the redundancy of the generated representations. To the best of our knowledge, few prior works have employed the information bottleneck principle to explore both minimal sufficient multimodal and unimodal representations, and this paper offers a solution to explore the feasibility of information bottleneck on multimodal learning tasks.

B. Information Bottleneck

Information bottleneck (IB) is originally proposed for signal processing, which aims to find a shorter version of the signal and meanwhile preserve the maximum information of the signal [8]. More recently, variational information bottleneck (VIB) [9] bridges the gap between IB and deep learning, which approximates the information bottleneck constraints and enables IB to be applicable with deep learning. Nowadays, due to its ability to learn minimal informative representations, IB and VIB have been widely employed in computer vision [38], reinforcement learning [39], and natural language processing [40]. Our paper is based on VIB, adapting and improving VIB for multimodal sentiment analysis and emotion recognition.

Among all the IB variants, the works that aim to adopt IB for multi-view learning are the most related to our work [41]–[45]. Lee et al. [44], Wu et al. [43],

and Zhu et al. [45] extend VIB for multi-view learning, which first constrain the information flow of each view using VIB, and then obtain the joint representation via the Product-of-Expert (PoE) [42]. Similarly, Wang et al. develop a deep multi-view IB theory, aiming to maximize the mutual information between the labels and the learned joint representation and minimize the mutual information between the learned representation of each view and the original representation [46]. The main differences between the proposed method and above multi-view IB methods are: 1) instead of only applying the IB principle on each modality, we develop E-MIB, L-MIB and C-MIB to study multimodal information bottleneck more comprehensively. Our methods manage to reduce the redundancy of the generated multimodal representation via the E-MIB, and filter out noisy information of each modality via the L-MIB. Moreover, C-MIB manages to combine the advantages of L-MIB and E-MIB to regularize both unimodal and multimodal representations. In a word, the proposed method simultaneously applies the IB principle to unimodal representation learning and complicated multimodal representation learning, which enables the learning of more discriminative and expressive features; 2) the proposed MIB variants are flexible as they can integrate any fusion mechanisms rather than only applying PoE as in previous multi-view IB methods [42]–[45]. In this way, more complicated and effective fusion strategy can be utilized to provide higher expressive power and explore more sufficient multimodal dynamics, which is shown to be important in our experiment (see Section IV-F). The compatibility of any fusion mechanisms is one of the advantages of the proposed method over multi-view IB methods. Notably, Ma et al. [47] propose a multimodal information bottleneck for image-to-speech synthesis. However, the multimodal information bottleneck in [47] is fundamentally different from our method, whose information bottleneck mainly refers to the designed model architecture instead of the IB principle.

III. Model Architecture

In this section, we introduce the pipeline of our MIB variants in detail, and the diagram of C-MIB is illustrated in Fig. 1. MIB consists of three unimodal learning networks and a multimodal fusion network, where the optimization of these networks is driven by the information bottleneck principle. The pipelines of the three MIB variants are illustrated in Algorithm 1, 2, and 3, respectively. The detailed introduction is shown in the following sections.

A. Notations and Task Definition

The downstream tasks in this paper are multimodal sentiment analysis and multimodal emotion recognition. The input to the model is an utterance [48], i.e., a segment of a video bounded by a sentence. Each utterance has three modalities, i.e., acoustic (a), visual (v), and language (l) modalities. The input sequences of acoustic, visual, and language modalities are denoted as $\boldsymbol{U_a} \in \mathbb{R}^{T_a \times d_a}$,
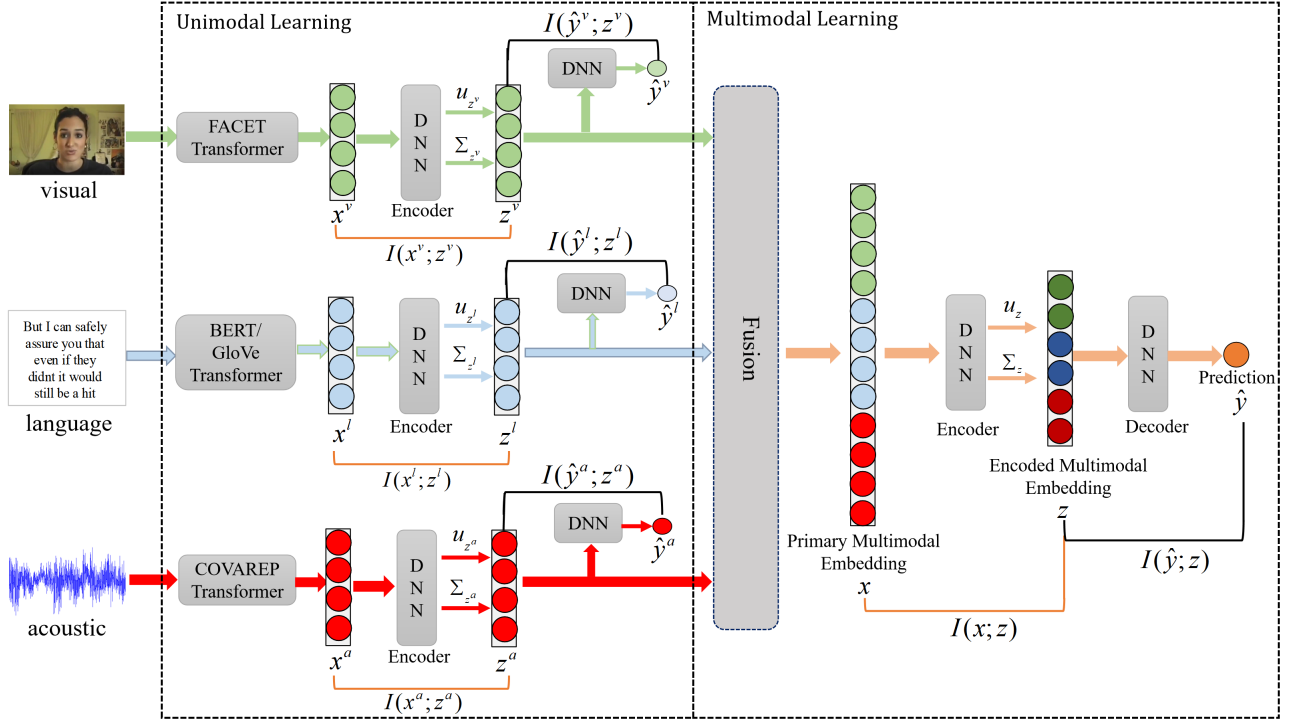
Fig. 1: The Diagram of C-MIB. DNN denotes deep neural network. The fusion mechanism in our framework is flexible.

Algorithm 1 Early-fusion MIB (E-MIB)

---

input: Unimodal representation $\boldsymbol{U}_m$, $m \in \{a, v, l\}$, hyper-parameter β

output: prediction $\hat{y}_i$

    Initialize unimodal learning networks F^m and multimodal fusion network F^f;

    while not done do

        Sample a batch of utterances

        for each utterance i do

            for each of modality m $(m \in \{l, a, v\})$ do

                $\boldsymbol{x}_i^m = \boldsymbol{F}^m(\boldsymbol{U}_i^m; \theta_m)$

            end for

            $\boldsymbol{x}_i = \boldsymbol{F}^f(\boldsymbol{x}_i^l, \boldsymbol{x}_i^a, \boldsymbol{x}_i^v; \theta_f)$

            $\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} = \boldsymbol{\mu}(\boldsymbol{x}_i; \theta_\mu), \boldsymbol{\Sigma}(\boldsymbol{x}_i; \theta_\Sigma)$

            $\boldsymbol{z}_i = \boldsymbol{\mu}_{z_i} + \boldsymbol{\Sigma}_{z_i} \times \boldsymbol{\epsilon}$

            $\hat{y}_i = \boldsymbol{D}(\boldsymbol{z}_i; \theta_d)$

        end for

        Compute J_{E-MIB} as in Eq. 15

    end while

Algorithm 2 Late-fusion MIB (L-MIB)

---

input: Unimodal representation $\boldsymbol{U}_m$, $m \in \{a, v, l\}$, hyper-parameter β

output: prediction $\hat{y}_i$

    Initialize unimodal learning networks F^m and multimodal fusion network F^f;

    while not done do

        Sample a batch of utterances

        for each utterance i do

            for each of modality m $(m \in \{l, a, v\})$ do

                $\boldsymbol{x}_i^m = \boldsymbol{F}^m(\boldsymbol{U}_i^m; \theta_m)$

                $\boldsymbol{\mu}_{z_i^m}, \boldsymbol{\Sigma}_{z_i^m} = \boldsymbol{\mu}^m(\boldsymbol{x}_i^m; \theta_\mu^m), \boldsymbol{\Sigma}^m(\boldsymbol{x}_i^m; \theta_\Sigma^m)$

                $\boldsymbol{z}_i^m = \boldsymbol{\mu}_{z_i^m} + \boldsymbol{\Sigma}_{z_i^m} \times \boldsymbol{\epsilon}$

                $\hat{y}_i^m = \boldsymbol{D^m}(\boldsymbol{z}_i^m; \theta_d^m)$

            end for

            $\boldsymbol{z}_i = \boldsymbol{F}^f(\boldsymbol{z}_i^l, \boldsymbol{z}_i^a, \boldsymbol{z}_i^v; \theta_f)$

            $\hat{y}_i = \boldsymbol{D}(\boldsymbol{z}_i; \theta_d)$

        end for

        Compute J_{L-MIB} as in Eq. 17

    end while

$\boldsymbol{U_v} \in \mathbb{R}^{T_v \times d_v}$, and $\boldsymbol{U_l} \in \mathbb{R}^{T_l \times d_l}$ respectively, where T_m and d_m denotes the sequence length and the feature dimensionality respectively ($m \in \{a, v, l\}$). Given the three unimodal sequences, we aim to learn a joint multimodal representation and then predict the sentiment score or emotion label based on the multimodal representation.

After being processed by unimodal learning network F^m, the input feature sequence $\boldsymbol{U_m} \in \mathbb{R}^{T_m \times d_m}$ becomes the unimodal representation $\boldsymbol{x_m} \in \mathbb{R}^{d_m}$. For conciseness, the structures of unimodal learning network F^m and the fusion network F^f are illustrated in the Appendix. Note that our algorithm is independent of the concrete fusion

mechanism, and we can inject various fusion methods into the multimodal fusion network to provide higher expressive power. In this paper, we mainly investigate five fusion methods to verify the effectiveness of our algorithm and comprehensively compare the performance between MIB variants.

B. Multimodal Information Bottleneck

In this section, we introduce the proposed MIB variants in detail. We extend the traditional information bottleneck

---

Algorithm 3 Complete MIB (C-MIB)

---

input: Unimodal representation $\boldsymbol{U}_m, \; m \in \{a, v, l\}$,
hyper-parameter β
output: prediction $\hat{y}_i$

    Initialize unimodal learning networks F^m and
    multimodal fusion network F^f;
    while not done do
        Sample a batch of utterances
        for each utterance i do
            for each of modality $m \;\; (m \in \{l, a, v\})$ do
                $\boldsymbol{x}_i^m = \boldsymbol{F}^m(\boldsymbol{U}_i^m; \theta_m)$
                $\boldsymbol{\mu}_{z_i^m}, \boldsymbol{\Sigma}_{z_i^m} = \boldsymbol{\mu^m}(\boldsymbol{x}_i^m; \theta_\mu^m), \boldsymbol{\Sigma^m}(\boldsymbol{x}_i^m; \theta_\Sigma^m)$
                $\boldsymbol{z}_i^m = \boldsymbol{\mu}_{z_i^m} + \boldsymbol{\Sigma}_{z_i^m} \times \boldsymbol{\epsilon}$
                $\hat{y}_i^m = \boldsymbol{D^m}(\boldsymbol{z}_i^m; \theta_d^m)$
            end for
            $\boldsymbol{x}_i = \boldsymbol{F}^f(\boldsymbol{z}_i^l, \boldsymbol{z}_i^a, \boldsymbol{z}_i^v; \theta_f)$
            $\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} = \boldsymbol{\mu}(\boldsymbol{x}_i; \theta_\mu), \boldsymbol{\Sigma}(\boldsymbol{x}_i; \theta_\Sigma)$
            $\boldsymbol{z}_i = \boldsymbol{\mu}_{z_i} + \boldsymbol{\Sigma}_{z_i} \times \boldsymbol{\epsilon}$
        end for
        Compute J_{C-MIB} as in Eqs. 21, 22, 23
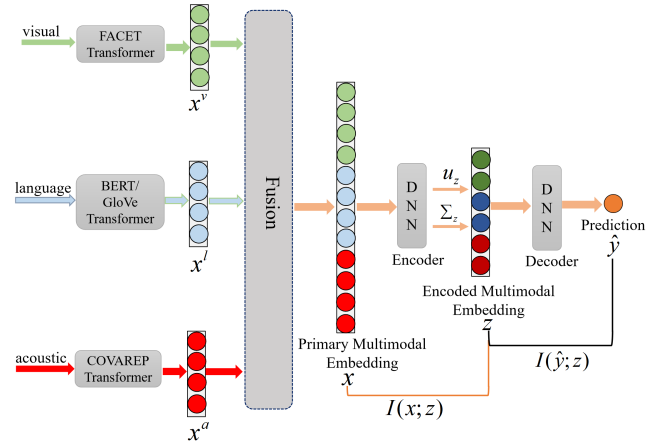   end while

---

Fig. 2: The Schematic Diagram of E-MIB.

imally discriminative about the target y (i.e., $I(y; \boldsymbol{z})$ is maximized). Obviously, the maximally informative representation can always be obtained by taking the identity mapping, i.e. $\boldsymbol{x} = \boldsymbol{z}$, which however contains noisy information or is often redundant for prediction. Therefore, a constraint is added on the MI between $\boldsymbol{z}$ and $\boldsymbol{x}$. The objective of IB thereby becomes:

$$(1) \; \text{maximize} \;\; I(y; \boldsymbol{z}) \quad \text{and} \quad (2) \; \text{minimize} \;\; I(\boldsymbol{x}; \boldsymbol{z}) \tag{2}$$

where the first constraint encourages $\boldsymbol{z}$ to be maximally predictive of the target y, and the second constraint forces $\boldsymbol{z}$ to contain as less information from $\boldsymbol{x}$ as possible. In other words, the IB principle aims to learn the minimal sufficient representation of $\boldsymbol{x}$ with respect to the label y, explicitly enforcing the $\boldsymbol{z}$ to only preserve the information in $\boldsymbol{x}$ that is discriminative to the prediction. In this paper, we refer the second constraint as the minimal information constraint. A common objective function for IB is defined as:

$$L_{IB} = I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}; \boldsymbol{z}) \tag{3}$$

where $\beta \geq 0$ is a scalar that determines the weight of the minimal information constraint during optimization.

2) Procedure of Early-Fusion MIB: Different from the traditional IB, in our multimodal system, the input contains three unimodal representations $\boldsymbol{x}^m$ output by the unimodal learning networks $\boldsymbol{F}^m \; (m \in \{a, v, l\})$. How to enforce the learned multimodal representation to contain as less information from the three unimodal representations as possible remains a problem. One intuitive way is to firstly fuse the unimodal representations into a joint multimodal representation, and then regularize the multimodal representation via IB, which is referred as early-fusion MIB (E-MIB). The pipeline of E-MIB is presented in Fig. 2. In this case, the objective function of E-MIB can be defined as:

$$L_{MIB} = I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}^a, \boldsymbol{x}^v, \boldsymbol{x}^l; \boldsymbol{z}) \tag{4}$$

where y is the label, $\boldsymbol{z}$ is the encoded multimodal representation, and $\boldsymbol{x}^a, \; \boldsymbol{x}^v, \; \boldsymbol{x}^l$ refer to the three unimodal representations output by the unimodal learning networks (see Appendix), respectively. To resolve the

(IB) to learn minimal sufficient multimodal representation, which contains the necessary information that is discriminative for the label and is free of noise. Moreover, we also add the IB constraint on the encoded unimodal representations to filter out the noisy information in unimodal representations that might interfere the learning of cross-modal interactions.

1) Information Bottleneck: Firstly, we introduce the general principle of information bottleneck (IB). IB aims to find better representation under the constraint on its complexity, such that the latent representation is discriminative and simultaneously free of redundancy. The IB principle defines what we mean by a good representation, in terms of the fundamental tradeoff between the concise representation and the good predictive power [9]. Specifically, IB is based on Mutual information (MI), aiming to maximize the MI between the encoded representation and the label as well as minimize the MI between the encoded representation and the input.

MI measures the amount of information obtained in one random variable after observing another random variable. Formally, given two random variables $\boldsymbol{x}$ and y with joint distribution $p(\boldsymbol{x}, y)$ and marginal distributions $p(\boldsymbol{x})$ and $p(y)$, their MI is defined as the KL-divergence between the joint distribution and the product of their marginal distributions. The equation is shown as follows:

$$\begin{aligned} I(\boldsymbol{x}; y) = I(y; \boldsymbol{x}) &= KL\Big(p(\boldsymbol{x}, y) || p(\boldsymbol{x})p(y)\Big) \\ &= \int d\boldsymbol{x} dy p(\boldsymbol{x}, y) \log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)} \\ &= \mathbb{E}_{(\boldsymbol{x}, y) \sim p(\boldsymbol{x}, y)} \left[\log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)} \right] \end{aligned} \tag{1}$$

The goal of IB is to learn the compressed encoded representation $\boldsymbol{z}$ using the input $\boldsymbol{x}$, where $\boldsymbol{z}$ is max-

problem of multiple input modalities, we firstly fuse the three modalities to obtain a primary multimodal representation $\boldsymbol{x}$ via the fusion mechanisms introduced in Appendix. Then we reduce the redundancy of the primary multimodal representation and filter out noisy information via the IB principle (that is why we call it early fusion). Specifically, the defaulted fusion mechanism is defined as the Concatenation: $\boldsymbol{x} = \boldsymbol{x}^a \oplus \boldsymbol{x}^v \oplus \boldsymbol{x}^l$. By the Concatenation, the $\boldsymbol{x}$ can preserve all the information from the three modalities, which is used to learn the encoded multimodal representation $\boldsymbol{z}$ that is minimal sufficient. The objective function thus can be rewritten as:

$$
\begin{aligned}
L_{E-MIB} &= I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}^a, \boldsymbol{x}^v, \boldsymbol{x}^l; \boldsymbol{z}) \\
&= I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}; \boldsymbol{z}) \quad (5) \\
&\text{s.t. } \boldsymbol{x} = F^f(\boldsymbol{x}^a, \boldsymbol{x}^v, \boldsymbol{x}^l; \ \theta_f)
\end{aligned}
$$

Note that our E-MIB differs from previous multi-view IB methods [43]–[45] in two main perspectives: 1) previous methods tend to apply the IB principle on each view. In contrast, the proposed E-MIB can reduce the redundant information in the generated complex multimodal embedding. Actually, the multimodal embedding integrates more information and is often high-dimensional and redundant, which is directly used for prediction. Therefore, applying the IB principle on multimodal representation is more reasonable and effective; 2) existing multi-view IB methods [43]–[45] mainly leverage PoE [42] to fuse information of each view to obtain the joint representation. Unlike them, our proposed E-MIB is flexible to integrate commonly-used fusion mechanisms in addition to PoE. In this way, more complicated and effective fusion strategies can be utilized to provide higher expressive power and explore more sufficient multimodal dynamics without redundancy. The compatibility of any commonly-used fusion mechanisms is one of the advantages of our MIB variants.

To optimize the objective function in Eq. 5, we adopt the solution introduced in VIB [9]. For brevity, we place the deduction in Appendix and only show the result of deduction:

$$
I(y; \boldsymbol{z}) \geq \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{z} \mid \boldsymbol{x}) p(y \mid \boldsymbol{x}) p(\boldsymbol{x}) \log q(y \mid \boldsymbol{z}) \quad (6)
$$

where $q(y \mid \boldsymbol{z})$ is a variational approximation to $p(y \mid \boldsymbol{z})$. The entropy of the target label, i.e., $H(y) = -\int dy p(y) \log p(y)$ is independent of the parameter optimization and thereby can be ignored. By this means, we can instead turn to maximize the lower bound of the objective function to optimize $I(y; \boldsymbol{z})$.

For the second term of the E-MIB (i.e., the minimal information constraint), we also place the detailed deduction in the Appendix and only present the result of deduction:

$$
I(\boldsymbol{x}; \boldsymbol{z}) \leq \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x}) p(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \quad (7)
$$

where $q(\boldsymbol{z})$ is a variational approximation to the marginal distribution $p(\boldsymbol{z})$ which is often fixed to a standard normal Gaussian distribution. Combining the above two constraints, we can obtain a lower bound of the objective function:

$$
\begin{aligned}
L_{E-MIB} &= I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}; \boldsymbol{z}) \geq J_{E-MIB} \\
&= \mathbb{E}_{(\boldsymbol{x},y) \sim p(\boldsymbol{x},y), \boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x})} \Big[\log q(y \mid \boldsymbol{z}) - \beta \cdot KL\big(p(\boldsymbol{z} \mid \boldsymbol{x}) \| q(\boldsymbol{z})\big) \Big]
\end{aligned}
$$
(8)

where J_{E-MIB} is a lower bound of L_{E-MIB}. By maximizing J_{E-MIB}, the lower bound of L_{E-MIB} is improved and thus L_{E-MIB} can be optimized.

To use deep neural networks to optimize J_{E-MIB}, we assume that $p(\boldsymbol{z} \mid \boldsymbol{x})$ is a Gaussian distribution. The mean and variance of the Gaussian distributions $p(\boldsymbol{z} \mid \boldsymbol{x})$ are learned using the respective deep neural network:

$$
p(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\Big(\boldsymbol{\mu}(\boldsymbol{x}; \theta_\mu), \boldsymbol{\Sigma}(\boldsymbol{x}; \theta_\Sigma)\Big) = \mathcal{N}\Big(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z\Big) \quad (9)
$$

where $\boldsymbol{\mu}$ parameterized by θ_μ and $\boldsymbol{\Sigma}$ parameterized by θ_Σ are the deep neural networks to learn the mean $\boldsymbol{\mu}_z$ and variance $\boldsymbol{\Sigma}_z$ of Gaussian distributions. However, in this way, the update of the parameters will be a problem because the computation of the gradients of parameters suffers from randomness. To resolve this problem, following [9], [46], we use the reparameterization trick to obtain $\boldsymbol{z}$:

$$
\boldsymbol{z} = \boldsymbol{\mu}_z + \boldsymbol{\Sigma}_z \times \boldsymbol{\epsilon} \quad (10)
$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ is a standard normal Gaussian distribution, and I is the identity vector whose elements are all equal to 1. By using the reparameterization trick, the randomness is transferred to $\boldsymbol{\epsilon}$ and the gradients of the parameters can be explicitly computed. Note that here we assume that each element in the vector $\boldsymbol{z}$ is independent from each other.

Moreover, for regression task, we formulate $q(y \mid \boldsymbol{z})$ as:

$$
\begin{aligned}
q(y \mid \boldsymbol{z}) &= e^{-||y - \boldsymbol{D}(\boldsymbol{z}; \theta_d)||_1 + C} \\
\log q(y \mid \boldsymbol{z}) &= -||y - \boldsymbol{D}(\boldsymbol{z}; \theta_d)||_1 + C = -||y - \hat{y}||_1 + C
\end{aligned}
$$
(11)

where C is a constant, $\boldsymbol{D}$ is a decoder parameterized by θ_d, and $\hat{y}$ is the prediction output by E-MIB. Under such circumstance, maximizing of $q(y \mid \boldsymbol{z})$ is equivalent to minimize the mean absolute error (MAE) between the prediction $\hat{y}$ and the target y [45]. And for classification task, $\log q(y \mid \boldsymbol{z})$ is formulated as:

$$
\begin{aligned}
q(y \mid \boldsymbol{z}) &= \hat{y}^y \cdot (1 - \hat{y})^{1-y}, \quad \hat{y} = \text{Sigmoid}(\boldsymbol{D}(\boldsymbol{z}; \theta_d)) \\
\log q(y \mid \boldsymbol{z}) &= y \log \hat{y} + (1 - y) log(1 - \hat{y})
\end{aligned}
$$
(12)

Here maximizing of $q(y \mid \boldsymbol{z})$ is equivalent to minimize the cross-entropy between the prediction $\hat{y}$ and the target y. In general, MAE and cross-entropy loss is used to maximize the mutual information between the target and the encoded multimodal representation $\boldsymbol{z}$ for regression task and classification task, respectively. Furthermore, in practice, the approximated marginal distribution of the encoded multimodal representation $\boldsymbol{z}$ is often assumed to be a standard normal Gaussian distribution [9], [46]:
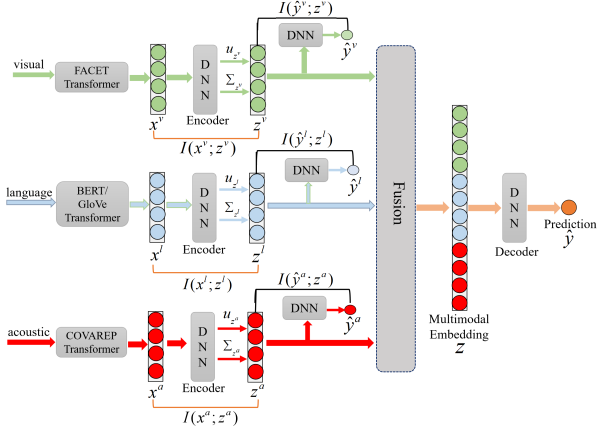
$$
q(\boldsymbol{z}) \sim \mathcal{N}(0, I) \quad (13)
$$

Fig. 3: The Schematic Diagram of L-MIB.

Combining Eq. 9 and Eq. 13, the KL-divergence term, i.e., $KL\big(p(\boldsymbol{z}\,|\,\boldsymbol{x})||q(\boldsymbol{z})\big)$ can be calculated as:

$$KL\big(p(\boldsymbol{z}\,|\,\boldsymbol{x})||q(\boldsymbol{z})\big) = KL\big(\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x};\theta_\mu),\boldsymbol{\Sigma}(\boldsymbol{x};\theta_\Sigma))||\mathcal{N}(0,I)\big)$$
$$= KL\big(\mathcal{N}(\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)||\mathcal{N}(0,I)\big) \tag{14}$$

Here we assume the choice of reparameterization of $p(\boldsymbol{z}\,|\,\boldsymbol{x})$ and $q(\boldsymbol{z})$ allows for the computation of an analytic KL-divergence. The KL-divergence of two Gaussian variables can be calculated by calculus and the computation is omitted here for conciseness.

Finally, the integral over $\boldsymbol{x}$, $\boldsymbol{z}$ and y can be approximated by Monte Carlo sampling [49], and thus J_{E-MIB} (Eq. 8) can be simplified as:

$$J_{E-MIB} \approx \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[\log q(y_i\,|\,\boldsymbol{z}_i)] - \beta\cdot KL\big(p(\boldsymbol{z}_i\,|\,\boldsymbol{x}_i)||q(\boldsymbol{z}_i)\big)\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[\log q(y_i\,|\,\boldsymbol{z}_i)] - \beta\cdot KL\big(\mathcal{N}(\boldsymbol{\mu}_{z_i},\boldsymbol{\Sigma}_{z_i})||\mathcal{N}(0,I)\big)\right] \tag{15}$$

where n is the sampling size (i.e., batch size), and i is the subscript that indicates each sample (which is omitted in the above equations for conciseness). The concrete form of $\log q(y_i\,|\,\boldsymbol{z}_i)$ depends on the task type (regression or classification). By maximizing J_{E-MIB}, we encourage $\boldsymbol{z}$ to be maximally discriminative to the target and simultaneously forget the information about the primary multimodal representation $\boldsymbol{x}$, by which means $\boldsymbol{z}$ ideally is no longer redundant.

3) Procedure of Late-Fusion MIB (L-MIB): A disadvantage of E-MIB is that unimodal features might contain noise that negatively affects the exploration of cross-modal interactions during fusion. Moreover, the distribution discrepancy of unimodal representations might seriously prevent the model from discovering complementary information between modalities. Therefore, we propose another version of MIB, namely late-fusion MIB (L-MIB), which allows unimodal representations to filter out noisy information before fusion and match the encoded unimodal distributions by individually applying the IB

principle on each modality. The pipeline of L-MIB is shown in Fig. 3, and the objective of L-MIB is:

$$L_{L-MIB} = I(y;\boldsymbol{z}) + \sum_m [I(y;\boldsymbol{z}^m) - \beta I(\boldsymbol{x}^m;\boldsymbol{z}^m)]$$
$$\text{s.t. } \boldsymbol{z} = F^f(\boldsymbol{z}^a,\boldsymbol{z}^v,\boldsymbol{z}^l;\theta_f) \tag{16}$$

where F^f is the multimodal fusion network described in the Appendix. From the deduction above, L_{L-MIB} can be approximated by the following objective:

$$J_{L-MIB} \approx \frac{1}{n}\sum_{i=1}^{n}\Bigg[\log q(y_i\,|\,\boldsymbol{z}_i)+$$
$$\sum_m\Big[\mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[\log q(y_i\,|\,\boldsymbol{z}_i^m)] - \beta\cdot KL\Big(p(\boldsymbol{z}_i^m\,|\,\boldsymbol{x}_i^m)||q(\boldsymbol{z}_i^m)\Big)\Big]\Bigg]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\log q(y_i\,|\,\boldsymbol{z}_i) + \sum_m\Big[\mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[\log q(y_i\,|\,\boldsymbol{z}_i^m)]$$
$$- \beta\cdot KL\Big(\mathcal{N}(\boldsymbol{\mu}_{z_i^m},\boldsymbol{\Sigma}_{z_i^m})||\mathcal{N}(0,I)\Big)\Big]\Bigg] \tag{17}$$

where,

$$p(\boldsymbol{z}^m\,|\,\boldsymbol{x}^m) = \mathcal{N}\big(\boldsymbol{\mu^m}(\boldsymbol{x}^m;\theta_\mu^m),\boldsymbol{\Sigma^m}(\boldsymbol{x^m};\theta_\Sigma^m)\big)$$
$$= \mathcal{N}\big(\boldsymbol{\mu}_{z^m},\boldsymbol{\Sigma}_{z^m}\big) \tag{18}$$
$$\boldsymbol{z}^m = \boldsymbol{\mu}_{z^m} + \boldsymbol{\Sigma}_{z^m}\times\boldsymbol{\epsilon}$$

where $\boldsymbol{D^m}$ is the decoder for modality m, $\boldsymbol{\mu^m}$ and $\boldsymbol{\Sigma^m}$ are the deep neural networks that learn the mean and the variance of $p(\boldsymbol{z}^m\,|\,\boldsymbol{x}^m)$ respectively. From the perspective of filtering out noisy information before fusion, L-MIB is effective as it forces the encoded unimodal representation $\boldsymbol{z}^m$ to only contain the relevant information to the target. Moreover, by using the reparameterization trick, the distributions of different encoded unimodal representations are closer, which narrows down the modality gap. The encoded unimodal representations are then fused to obtain the final multimodal representation $\boldsymbol{z}$, where $\boldsymbol{z}$ is used to infer the final prediction.

Similarly, the fusion procedure is flexible and we can integrate any fusion methods into L-MIB to obtain the encoded multimodal representation $\boldsymbol{z}$. Actually, another way to obtain $\boldsymbol{z}$ for L-MIB is referred to as the Bayesian product-of-experts (PoE) system [42], [43], [46]:

$$\boldsymbol{\mu}_z = (\sum_m \boldsymbol{\mu}_{z^m}\cdot\boldsymbol{\Sigma}_{z^m}^{-1})(I + \sum_m \boldsymbol{\Sigma}_{z^m}^{-1})^{-1}$$
$$\boldsymbol{\Sigma}_z = (I + \sum_m \boldsymbol{\Sigma}_{z^m}^{-1})^{-1} \tag{19}$$
$$\boldsymbol{z} = \boldsymbol{\mu}_z + \boldsymbol{\Sigma}_z\times\boldsymbol{\epsilon}$$

where $\cdot$ denotes element-wise multiplication. PoE is similar to weighted average. The mean part $\boldsymbol{\mu}_z$ of the $\boldsymbol{z}$ is actually the average of modality-specific $\boldsymbol{\mu}_{z^m}$ weighted by the reciprocal of the corresponding variance $\boldsymbol{\Sigma}_{z^m}$. PoE assumes the experts with greater confidence are the modalities whose information bears less uncertainty and is more relevant to the target. For experts with greater confidence (i.e., lower variance), they will have larger weights to obtain $\boldsymbol{z}$ than those experts with higher

variance. PoE severs as a fusion baseline and we will compare it with other fusion methods in Section IV-F.

Incorporating the IB principle for each modality helps to filter out the noise and focus on the discrimiative unimodal features. Especially for the acoustic and visual modalities, they contain much noisy information that interferes the prediction [3], which will negatively affect the quality of the fused multimodal representation (especially when complex fusion methods are applied). Therefore, it is beneficial to filter out the noise in the representations of these modalities. However, by applying the IB principle on unimodal branches, the concurrence of the information or the complementary information between multiple modalities (which is not necessarily discriminative) might be ignored. Nevertheless, we argue that the hyperparameter β in the proposed framework can alleviate this issue to some extent, balancing the non-discriminative and discriminative information existing in different modalities.

4) Procedure of C-MIB: C-MIB is built to combine the advantages of E-MIB and L-MIB. Firstly, C-MIB applies the IB principle to extract encoded unimodal representations that ideally are free of noise. Then the encoded unimodal representations are fused to obtain the primary multimodal representation $\boldsymbol{x}$. The IB principle is further applied on $\boldsymbol{x}$ to reduce the redundancy and obtain the final multimodal representation $\boldsymbol{z}$. The objective of C-MIB is shown in:

$$
\begin{aligned}
L_{C-MIB} =& I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}; \boldsymbol{z}) + \\
& \sum_m [I(y; \boldsymbol{z}^m) - \beta I(\boldsymbol{x}^m; \boldsymbol{z}^m)] \\
& \text{s.t. } \boldsymbol{x} = F^f(\boldsymbol{z}^a, \boldsymbol{z}^v, \boldsymbol{z}^l; \theta_M)
\end{aligned}
\tag{20}
$$

From the deduction above, L_{C-MIB} can be approximated using J_{C-MIB}:

$$
J_{C-MIB} \approx \frac{1}{n} \sum_{i=1}^{n} \left[J_i^M + \sum_m J_i^m \right]
\tag{21}
$$

where,

$$
\begin{aligned}
J_i^M &= \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})}[\log q(y_i \,|\, \boldsymbol{z}_i)] - \beta \cdot KL\Big(p(\boldsymbol{z}_i \,|\, \boldsymbol{x}_i) || q(\boldsymbol{z}_i)\Big) \\
&= \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})}[\log q(y_i \,|\, \boldsymbol{z}_i)] - \beta \cdot KL\Big(\mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) || \mathcal{N}(0, I)\Big)
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
J_i^m &= \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})}[\log q(y_i \,|\, \boldsymbol{z}_i^m)] - \beta \cdot KL\Big(p(\boldsymbol{z}_i^m \,|\, \boldsymbol{x}_i^m) || q(\boldsymbol{z}_i^m)\Big) \\
&= \mathbb{E}_{\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})}[\log q(y_i \,|\, \boldsymbol{z}_i^m)] - \beta \cdot KL\Big(\mathcal{N}(\boldsymbol{\mu}_{z_i^m}, \boldsymbol{\Sigma}_{z_i^m}) || \mathcal{N}(0, I)\Big)
\end{aligned}
\tag{23}
$$

The comparison of E-MIB, L-MIB and C-MIB is shown in experiment section.

IV. Experiments

In this section, we mainly evaluate our model on the multimodal sentiment analysis task. To evaluate the generalization ability of our model to other multimodal learning tasks, we additionally present the results of our model on the multimodal emotion recognition task.

TABLE I: Comparison with baselines on CMU-MOSI. 'B-MIB' refers to the baseline MIB model whose objective is maximizing the mutual information between multimodal representation and the label, which is implemented by removing the minimal information constraint from E-MIB.

	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
EF-LSTM	31.6	75.8	75.6	1.053	0.613
LF-LSTM	31.6	76.4	75.4	1.037	0.620
TFN [4]	32.2	76.4	76.3	1.017	0.604
LMF [22]	30.6	73.8	73.7	1.026	0.602
MFN [52]	32.1	78.0	76.0	1.010	0.635
RAVEN [30]	33.8	78.8	76.9	0.968	0.667
MULT [6]	33.6	79.3	78.3	1.009	0.667
QMF [32]	35.5	79.7	79.6	0.915	0.696
MISA [53]	42.3	83.4	83.6	0.783	0.761
MAG-BERT [35]	42.9	83.5	83.5	0.790	0.769
TFR-Net [54]	42.6	84.0	83.9	0.787	0.788
B-MIB	46.6	82.8	82.8	0.737	0.788
E-MIB	48.6	85.3	85.3	0.711	0.798
L-MIB	45.8	84.6	84.6	0.732	0.790
C-MIB	48.2	85.2	85.2	0.728	0.793

A. Datasets

1) CMU-MOSI: CMU-MOSI [50] is a widely-used dataset for multimodal sentiment analysis. It contains 93 videos in total, and each video is divided into 62 utterances at most. The intensity of sentiment of each utterance is within [-3,3], where -3 indicates the strongest negative sentiment, and +3 the strongest positive.

2) CMU-MOSEI: CMU-MOSEI [26] is a large multimodal sentiment analysis and emotion recognition dataset that contains a total number of 2928 videos. The dataset has been segmented at the utterance level, and each utterance has been scored on two levels: sentiment ranging between [-3, 3], and emotion with six different values (i.e., anger, disgust, fear, happy, sad, and surprise). We predict the sentiment and emotion score of the utterance for this dataset.

3) IEMOCAP: IEMOCAP [51] is a multimodal emotion recognition dataset that contains a total number of 151 videos from 10 speakers. The videos are segmented into about 10K utterances. The dataset has the following labels: anger, happiness, sadness, neutral, excitement, frustration, fear, surprise and other. We take the first four emotions to compare with our baselines. We follow previous works [3], [6] to report the classification accuracy and the F1 score of each emotion.

B. Experimental Details

We develop our model using PyTorch on GTX2080Ti with CUDA 10.1 and torch 1.4.0. The defaulted fusion method is set to Concatenation for all the MIB variants. More experimental details, evaluation protocol, and the introduction of baselines are shown in the Appendix for lack of space.

C. Comparison on Multimodal Sentiment Analysis

In this section, we compare our proposed MIB algorithms with our baselines on the multimodal sentiment analysis task. The results are shown in Table I and II, where the best results are highlighted in bold and the

TABLE II: Comparison with baselines on CMU-MOSEI.

	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
EF-LSTM	46.7	79.1	78.8	0.665	0.621
LF-LSTM	49.1	79.4	80.0	0.625	0.655
TFN [4]	49.8	79.4	79.7	0.610	0.671
LMF [22]	50.0	80.6	81.0	0.608	0.677
MFN [52]	49.1	79.6	80.6	0.618	0.670
RAVEN [30]	50.2	79.0	79.4	0.605	0.680
MULT [6]	48.2	80.2	80.5	0.638	0.659
IMR [33]	48.7	80.6	81.0	-	-
QMF [32]	47.9	80.7	79.8	0.640	0.658
MISA [53]	52.2	85.5	85.3	0.555	0.756
MAG-BERT [35]	51.9	85.0	85.0	0.602	0.778
TFR-Net [54]	51.7	85.2	85.1	0.606	0.781
B-MIB	52.1	84.4	84.3	0.603	0.775
E-MIB	53.2	85.4	85.4	0.588	0.790
L-MIB	54.1	85.7	85.7	0.600	0.788
C-MIB	53.0	86.2	86.2	0.584	0.789

TABLE III: Comparison with baselines on IEMOCAP.

Models	Happy		Sad		Angry		Neutral	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MFN [52]	86.5	84.0	83.5	82.1	85.0	83.7	69.6	69.2
Graph-MFN [26]	86.8	84.2	83.8	83.0	85.8	85.5	69.4	68.9
RAVEN [30]	87.3	85.8	83.4	83.1	87.3	86.7	69.7	69.3
LMF [22]	86.9	82.3	85.4	84.7	87.1	86.8	71.6	71.4
MULT [6]	87.4	84.1	84.2	83.1	88.0	87.5	69.9	68.4
HFFN [7]	86.8	82.1	84.4	84.5	86.6	85.8	69.6	69.3
TCM-LSTM [3]	87.2	84.8	84.4	84.9	89.0	88.6	71.3	71.2
E-MIB	88.1	85.9	86.8	86.9	88.8	88.7	71.0	70.9
L-MIB	88.4	86.5	86.9	86.1	88.3	88.3	70.9	70.6
C-MIB	87.7	85.8	87.6	86.9	89.2	88.8	72.4	72.1

variants become consistent when various fusion methods are considered.

second best results are marked with underlines. For CMU-MOSI dataset, although TFR-Net [54] performs better than the existing methods and sets up a high baseline, our MIB variants still obtain the best performance on the majority of the evaluation metrics. Specifically, E-MIB achieves the best results on all metrics, and it outperforms TFR-Net [54] by 6.0% on Acc7, 1.3% on Acc2 and 1.4% on F1 score. On CMU-MOSEI dataset, our proposed C-MIB yields 0.7% improvement on Acc2 and 1.2% improvement on F1 score compared to the current state-of-the-art MISA [53]. The performance on Acc7 and Corr is also improved compared to the state-of-the-art methods. These results demonstrate the effectiveness of our MIB variants, indicating the importance of learning minimal sufficient representations.

Notably, to compare with baselines, we only present the results of our MIB variants with the fusion method being Concatenation. Interestingly, combining the performance on the two datasets, E-MIB performs comparably to C-MIB. This is reasonable because when Concatenation is applied, the unimodal dynamics cannot fully interact with each other, and thus the noise in unimodal representations will not seriously influence the quality of the multimodal representation. Therefore, there is no significant gain when the IB principle is applied on the unimodal representations. Actually, although our model is agnostic of fusion architectures, the specific fusion method can still differently influence the performance of different MIB variants. Therefore, we propose to compare the performance between the MIB variants by comprehensively estimating their performance under different fusion strategies. For the comprehensive comparison between the MIB variants, please refer to Section IV-F. Also notice that the performance difference between the two different datasets exists. For instance, L-MIB performs weaker than E-MIB on CMU-MOSI, but it manages to perform better than E-MIB on CMU-MOSEI. The performance difference may be partly due to the reason that our methods introduce randomness when generating the latent representations, which might affect the performance, especially on a smaller dataset (i.e. CMU-MOSI). We also refer the reader to Section IV-F for the comprehensive comparison between the MIB variants, where the average performances of the MIB

D. Results on Multimodal Emotion Recognition

We additionally evaluate the MIB variants on the task of multimodal emotion recognition to justify the generalization ability of the models to other multimodal task. Two widely-used datasets, i.e. CMU-MOSEI and IEMOCAP are evaluated in this section. For IEMOCAP, as shown in Table III, our C-MIB achieves the best performance in the task of recognizing the 'Angry', 'Sad', and 'Neutral' emotions. Moreover, the L-MIB reaches the best performance of recognizing the 'Happy' emotion. The E-MIB achieves the best result on the F1 score of recognizing the 'Sad' emotion, yielding 2.0% improvement compared with the best result of existing methods. Notably, combining the performance on all the emotions, all the MIB variants outperform the baselines. For CMU-MOSEI, we can infer from Table IV that all MIB variants obtain very competitive performance on the average weighted acc and auc metrics. Our methods also outperform existing methods on the recognition of the majority of emotions. The extra experiments of multimodal emotion recognition have proven the effectiveness and generalization ability of our methods.

E. Ablation Study

In this section, we perform ablation studies on the C-MIB to investigate the contribution of each unimodal minimal information constraint loss by removing it from each modality, and the results are presented in Table V. We design three ablation experiments to investigate the contribution of each component in our proposed C-MIB. Firstly, we remove the IB principle from language modality (see the case of 'W/O IB on Language' where the language representation is directly used to generate multimodal representation $\boldsymbol{x}$), and the performance of the model has a slight drop. Secondly, we remove the IB principle on acoustic and visual modality respectively, and significant drop on performance is observed. These results demonstrate that the IB principle is more effective on visual and acoustic modalities, mainly for the reason that they have many redundant and noisy information that is irrelevant to the prediction. Since the language modality is far more discriminative [3], [7], [24], the influence is not that significant when the IB principle is removed from the language modality.

TABLE IV: Comparison with baselines on CMU-MOSEI dataset (multimodal emotion recognition task).

Models	Anger		Disgust		Fear		Happy		Sad		Surprise		Average	
	W-Acc	AUC	W-Acc	AUC	W-Acc	AUC	W-Acc	AUC	W-Acc	AUC	W-Acc	AUC	W-Acc	AUC
EF-LSTM	58.5	62.2	59.9	63.9	50.1	69.8	65.1	68.9	55.1	58.6	50.6	54.3	56.6	63.0
LF-LSTM	57.7	66.5	61.0	71.9	50.7	61.1	63.9	68.6	54.3	59.6	51.4	61.5	56.5	64.9
Graph-MFN [26]	62.6	-	69.1	-	62.0	-	66.3	-	60.4	-	53.7	-	62.3	-
MTL [28]	66.8	68.0	72.7	76.7	62.2	42.9	53.6	71.4	61.4	57.6	60.6	65.1	62.8	63.6
MTEE [55]	67.0	71.7	72.5	78.3	65.4	71.6	67.9	73.9	62.6	66.7	62.1	66.4	66.2	71.2
E-MIB	68.4	71.6	77.5	79.0	70.6	69.1	55.3	74.7	61.8	67.9	66.3	68.0	66.7	71.7
L-MIB	66.9	71.5	76.5	78.2	69.5	71.3	55.2	74.6	64.4	66.7	66.7	66.4	66.5	71.4
C-MIB	65.2	72.5	75.3	79.8	68.5	71.5	55.2	75.2	61.7	68.7	65.4	67.7	65.9	72.5

TABLE V: Ablation studies on the CMU-MOSI dataset.

	Acc7 (↑)	Acc2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)
W/O IB on Language	47.4	84.4	84.4	0.725	0.794
W/O IB on Acoustic	47.2	84.0	84.0	0.725	0.788
W/O IB on Visual	46.4	84.0	84.0	0.732	0.791
C-MIB	48.2	85.2	85.2	0.728	0.793

F. Discussion on Fusion Architectures and the MIB Variants

In this section, we comprehensively evaluate and compare the performance of the MIB variants by incorporating various fusion strategies. Moreover, we verify that the our MIB variants are compatible with various fusion strategies and can provide higher expressive power. Analyzing the results on Table VI and VII, we have the following observations:

1) The comparison between the MIB variants: We use the average results of all the fusion methods as the metrics for a more comprehensive comparison between the MIB variants. Specifically, C-MIB reaches the best performance on the majority of the average metrics on both datasets, while E-MIB reaches the best score on some of the average metrics. Combining all the metrics, C-MIB slightly outperforms E-MIB. L-MIB is the worst version among all the MIB variants. This is reasonable because the multimodal representation is directly used for prediction, and one should directly add the minimal information constraint on the multimodal representation to obtain a better performance. However, L-MIB only adds the constraints to the unimodal representations such that the generated multimodal representation might still be redundant (especially when high-dimensional multimodal representation is generated). Therefore, L-MIB only achieves a sub-optimal solution. Notably, the differences in the average performance between the MIB variants are consistent across two datasets.

Interestingly, we observe that when the complex fusion methods (tensor fusion and graph fusion) are applied, the performance of the C-MIB improves considerably. In contrast, the performances of the L-MIB and the E-MIB are not significantly improved. This is because E-MIB cannot explicitly filter out the noise in unimodal representations, and the noise may seriously influence the quality of the multimodal representation when complex fusion mechanisms are applied. For the L-MIB, it cannot reduce the redundancy of the multimodal representation. Actually, when complex fusion mechanisms are used, the generated multimodal representation is often highly complicated and redundant. Therefore, the performance of L-MIB is inferior to C-MIB.

2) The comparison with PoE: By comparing the performance of PoE [42], we demonstrate the advantage of our MIB over the IB methods applying on multi-view learning [43]–[45]. PoE [42] is similar to weighted average and is widely-used in many multi-view IB methods [42]–[45]. However, it performs unfavorably compared to other fusion methods injecting into L-MIB. It demonstrates that we should enable the framework to be compatible with other fusion methods such as tensor fusion to provide higher expressive power. The compatibility with other fusion methods is one of the advantages of our MIB framework but cannot be realized by previous multi-view IB methods [42]–[45].

3) The comparison between fusion methods: Concatenation, Addition, and Multiplication are the simple fusion methods, which cannot explore complex inter-modal dynamics. Compared with them, more complicated fusion strategies such as graph fusion and tensor fusion can further learn complex cross-modal dynamics that benefit prediction. However, they face a higher risk of generating redundant information and increasing the computational cost. In general, Concatenation outperforms the other simple fusion methods on most of the evaluation metrics across two datasets and three MIB variants. This is reasonable because Concatenation can well-preserve the unimodal information when generating the multimodal representation, which cannot be realized by Addition and Multiplication. In addition, combining the results from the two datasets, tensor fusion reaches the best performance among all the fusion methods, which slightly outperforms graph fusion. We argue that this is because tensor fusion can generate high-expressive and high-dimensional multimodal tensor to explore the complicated interactions between modalities. And via the minimal information constraint of our MIB, the redundancy and the noise of unimodal representations and the multimodal tensor are reduced, such that the generated multimodal tensor is of high quality and suitable for prediction.

G. Analysis on Unimodal and Bimodal Systems

In this section, we provide the results of the unimodal and bimodal systems as well as their counterparts where the IB regularization is removed. By observing the results in Table VIII, we have the following conclusions:

1) The importance of each modality: The language modality, with the powerful BERT [34] framework as its feature extractor, significantly performs better than the visual and acoustic modalities. Moreover, visual modality

TABLE VI: Discussion on the fusion strategies and the MIB variants on CMU-MOSI dataset. The best average results across all the fusion methods are highlighted, and the second best average results are marked with underlines.

	E-MIB					L-MIB					C-MIB				
	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
Concatenation	48.6	85.3	85.3	0.711	0.798	45.8	84.6	84.6	0.732	0.790	48.2	85.2	85.2	0.728	0.793
Addition	47.3	84.3	84.3	0.714	0.799	45.6	84.4	84.4	0.722	0.793	48.2	84.6	84.6	0.721	0.796
Multiplication	46.7	85.0	85.0	0.716	0.801	47.0	84.7	84.7	0.736	0.787	47.9	84.9	84.9	0.726	0.794
PoE	-	-	-	-	-	47.2	84.3	84.3	0.749	0.786	-	-	-	-	-
Tensor Fusion [4]	46.9	85.0	85.0	0.727	0.798	46.7	85.0	85.0	0.714	0.795	46.7	86.0	85.9	0.713	0.799
Graph Fusion [25]	45.8	85.2	85.2	0.726	0.783	46.9	84.3	84.3	0.732	0.791	47.1	85.5	85.4	0.716	0.794
Average	47.1	85.0	85.0	0.719	0.796	46.5	84.6	84.6	0.731	0.790	47.6	85.2	85.2	0.721	0.795

TABLE VII: Discussion on the fusion strategies and the MIB variants on CMU-MOSEI dataset.

	E-MIB					L-MIB					C-MIB				
	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
Concatenation	53.2	85.4	85.4	0.588	0.790	54.1	85.7	85.7	0.600	0.788	53.0	86.2	86.2	0.584	0.789
Addition	54.8	85.5	85.5	0.600	0.784	53.9	85.3	85.3	0.592	0.788	54.0	85.7	85.7	0.600	0.782
Multiplication	52.6	85.4	85.4	0.606	0.774	52.8	85.1	85.2	0.612	0.775	53.0	85.5	85.4	0.596	0.786
PoE	-	-	-	-	-	53.2	85.2	85.3	0.601	0.782	-	-	-	-	-
Tensor Fusion [4]	53.7	85.4	85.5	0.605	0.776	53.6	85.3	85.3	0.604	0.777	54.3	86.3	86.4	0.597	0.787
Graph Fusion [25]	54.6	85.5	85.3	0.590	0.786	53.5	85.4	85.4	0.598	0.779	54.5	86.3	86.2	0.590	0.790
Average	53.8	85.4	85.4	0.598	0.782	53.5	85.3	85.4	0.601	0.782	53.8	86.0	86.0	0.593	0.787

TABLE VIII: Discussion of unimodal and bimodal systems on the CMU-MOSI dataset. 'L', 'A', and 'V' denote language, acoustic, and visual modality, respectively.

	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
L	46.6	83.5	83.5	0.728	0.792
L (W/O IB)	46.0	83.1	83.0	0.731	0.795
A	18.4	55.7	54.2	1.478	0.126
A (W/O IB)	16.2	55.1	53.7	1.474	0.007
V	21.8	57.1	57.0	1.460	0.060
V (W/O IB)	18.1	53.9	54.2	1.477	0.059
L & A	45.7	84.4	84.4	0.713	0.801
L & A (W/O IB)	46.3	83.4	83.4	0.734	0.794
A & V	20.3	58.6	57.4	1.445	0.165
A & V (W/O IB)	18.7	54.1	53.9	1.490	0.025
L & V	46.6	84.3	84.3	0.717	0.800
L & V (W/O IB)	47.1	83.8	83.8	0.727	0.791

TABLE IX: Analysis on the Hyper-parameter β.

	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
$\beta =$1e0	38.0	80.5	80.5	0.872	0.724
$\beta =$1e-1	47.2	84.4	84.4	0.718	0.797
$\beta =$1e-2	46.7	84.9	84.9	0.731	0.793
$\beta =$1e-3	48.2	85.2	85.2	0.728	0.793
$\beta =$1e-4	47.6	84.7	84.7	0.728	0.793
$\beta =$1e-5	48.0	85.7	85.6	0.709	0.799
$\beta =$1e-6	46.6	84.3	84.3	0.727	0.795
$\beta =$0	46.3	83.8	83.8	0.730	0.790

performs better than acoustic modality. However, the performances of these two modalities are actually unsatisfactory. One of the reasons is that the extracted features of these two modalities are of low quality and not sentiment-sensitive. This highlights a possible direction for improving the performance of the multimodal sentiment analysis system: extracting high-quality features of the nonverbal modalities for sentiment analysis.

2) The discussion on bimodal systems: Generally, the bimodal systems outperform their unimodal counterparts, and the trimodal system performs better than its bimodal counterparts (see Table I for the results of trimodal system). These results demonstrate that integrating more modalities helps improve the performance. Notably, even with a lowest performance on acoustic modality, the language-acoustic system achieves the best performance among the bimodal systems. The result indicates that language and acoustic modalities have more complementary information between each other.

3) The effectiveness of the IB principle on unimodal and bimodal systems: As shown in Table VIII, adding the IB constraints consistently improves the overall performance. Moreover, we observe that the IB principle is more beneficial to those modalities with more noises: the improvement of the IB regularization is more significant on visual, acoustic, and visual-acoustic systems. These results to some extent demonstrate that the IB regularization helps reduce the noise in the features and improve the performance.

H. Analysis on the Hyperparameter β

In this section, we conduct analysis on the hyperparameter β which controls the strength of the minimal information constraint (see Eq. 20). β also balances the non-discriminative and discriminative information in the latent representations. The experiment is conducted on C-MIB. From the results in Table IX, we can observe that when β becomes larger (β=1), the performance drops significantly. Notably, when β is equal to zero, binary accuracy drops by 1.5% and 7-class accuracy drops by 2% compared to the performance when β is 1e-3, and the MAE and Corr also become worse. Actually, in the case that β is zero, the minimal information constraint is removed, and the framework degenerates into a regular framework whose objective is maximizing the mutual information between representations and the target. The decline on performance demonstrates that the introduced IB principle is effective and can boost the performance. We argue that this is because the minimal information constraint helps to filter out the noise and reduce the redundancy of unimodal and multimodal representations.

Actually, the model performs best when β is set to 1e-5. To be consistent with the setting of L-MIB and E-MIB, the defaulted value of β is set to 1e-3. All in all, these results indicate that β should be a small number such that the minimal information constraint will not dominate the learning and more information can be retained, which prevents the IB principle from filtering out the necessary information. Meanwhile, it is necessary to reduce the redundancy and filter out the noise, as the performance drops considerably when the minimal information constraint is removed.
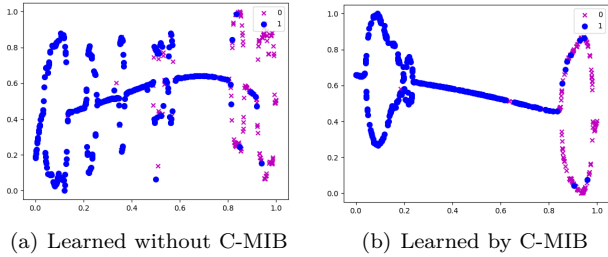
(a) Learned without C-MIB   (b) Learned by C-MIB

Fig. 4: Visualization of the multimodal representations on CMU-MOSI testing set. The 'purple x' and 'blue dot' denote the data point for positive and negative sentiment, respectively.

I. Visualization of the Embedding Space

In this section, we provide a visualization for distributions of the multimodal representations to analyze the class distributions in the embedding space. Following [25], [53], the visualization is obtained by transforming each representation into 2-dimensional feature point using the t-SNE algorithm [56], which can capture much of the local structure of the high-dimensional data very well. The left sub-figure on Fig. 4 illustrates the embedding space learnt without C-MIB, and the right sub-figure presents the embedding space learnt by the C-MIB. We can infer from Fig. 4 that when the IB principle is not applied, the data points of the multimodal representations in the embedding space are very scattered, and each class (positive or negative class) does not form a distinguishing cluster. In contrast, when the IB principle is applied, the data points from the same class form a discriminative cluster. The distance between the centers of the positive and negative sentiment clusters is far away, and the data points that are difficult to be clustered are placed between the two clusters. The visualization results reveal an important property of our C-MIB, i.e. enabling the multimodal representation to be more discrimiative by reducing the redundancy and filtering out the noise in unimodal and multimodal features.

J. Discussion on Minimal Information Constraint

The minimal information constraint aims to constrain the mutual information of the encoded representation z and the input x. It encourages z to forget the information from x. In this sense, the minimal information constraint can be replaced by other functions that control the 'similarity' between z and x. In this section, we replace the minimal information constraint (KL-divergence term) with other objectives to reveal its significance. The compared objectives include maximum mean discrepancy (MMD), Deep CORAL [57], orthogonal constraint, European distance and cosine distance. MMD is widely used in domain adaption to measure the distribution discrepancy between the source and target domain. Here we apply MMD to encourage the distribution discrepancy between z and x to be maximum such that they are not similar in terms of distribution. We use the Gaussian kernel as the kernel function of MMD. Since MMD is widely used,

TABLE X: Discussion on Minimal Information Constraint.

	Acc7 ($\uparrow$)	Acc2 ($\uparrow$)	F1 ($\uparrow$)	MAE ($\downarrow$)	Corr ($\uparrow$)
MMD	46.6	84.6	84.6	0.725	0.798
Deep CORAL	44.7	84.4	84.4	0.735	0.798
Orthogonal Constraint	47.6	84.9	84.9	0.717	0.798
European distance	45.8	84.6	84.6	0.732	0.795
Cosine Distance	46.6	83.8	83.8	0.719	0.800
C-MIB	48.2	85.2	85.2	0.728	0.793

we omit the computation of the MMD loss l'_{MDD} for conciseness. Since we need to maximize MMD, the final MMD loss is calculated as: $l_{MDD} = \frac{1}{1+l'_{MDD}}$ where 1 is used in the denominator to prevent division by zero. Deep CORAL is similar to MMD which adds constraints on the distribution between z and x:

$$C_x = \frac{1}{n-1}\left(x^T x - \frac{1}{n}\left(\mathbf{1}^T x\right)^T \left(\mathbf{1}^T x\right)\right)$$

$$C_z = \frac{1}{n-1}\left(z^T z - \frac{1}{n}\left(\mathbf{1}^T z\right)^T \left(\mathbf{1}^T z\right)\right) \quad (24)$$

$$l'_{CORAL} = \frac{1}{4d^2}\|C_S - C_T\|_F^2 \quad l_{CORAL} = \frac{1}{1+l'_{CORAL}}$$

where $\mathbf{1}$ is the column vector with all elements equal to 1, $\|\cdot\|_F^2$ denotes Frobenius norm, d is the feature dimensionality, n is the sampling size, and T denotes the transpose operation. In comparison, European distance forces the distance between z and x to be maximum such that they are not similar in terms of representations, and cosine distance likewise explains. Orthogonal constraint [53] forces z and x to be orthogonal:

$$x \leftarrow \frac{x}{\sqrt{\sum_j x_j^2} + 1}, \quad z \leftarrow \frac{z}{\sqrt{\sum_j z_j^2} + 1} \quad (25)$$

$$l_{oc} = x z^T$$

As presented in Table X, the original KL-divergence term performs best, and the other methods also reach competitive performance compared to the case that minimal information constraint is removed (see 'W/O IB' case on Table I). These results demonstrate that the idea of controlling the information from the input and filtering out noisy information is effective. Since the KL-divergence constraint explicitly measures the mutual information between the input and the encoded representation, it can reach a better performance compared to other methods that only measure the similarity of distribution or feature embedding between the input and the encoded one.

K. Learning Curve for Minimal Information Constraints

In this section, we present the learning curves for the minimal information constraint losses (i.e., the KL-divergence losses). As shown in Fig. 5, the losses of acoustic and visual modalities become much lower as the training deepens, indicating that these two modalities contain many noises that need to be filtered out. In contrast, the loss for language modality drops very little, suggesting that language modality contains sufficient discriminative information for predicting the target. These results are consistent with previous work [3], [7], [24], and
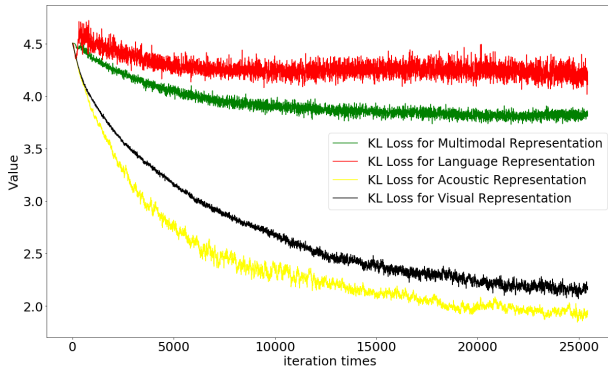
Fig. 5: Learning Curve for Minimal Information Constraint Losses of C-MIB. The 'KL Loss' in the figure denotes minimal information constraint loss.

indicate that our method can explicitly filter out noisy information and retain useful information. As for the loss for multimodal representation, it drops moderately and is convergent in the middle of the training, which suggests that the generated multimodal representation still has redundant information and needs to be explicitly handled.

TABLE XI: The comparison of model complexity.

	The number of parameters
MAG-BERT [35]	110,853,121
Ours (W/O IB)	109,216,699
E-MIB	109,371,423
L-MIB	109,681,024
C-MIB	109,835,748

L. Analysis on Model Complexity

We analyze the model complexity of the proposed method on CMU-MOSI in this section. Since our MIB variants require additional encoders and decoders to process the signal, the number of parameters increases after the injection of the IB regularization. We can infer from Table XI that when the IB principle is introduced, the number of parameters for the C-MIB is 109,835,748. The number of parameters becomes 109,216,699 when the IB constraints are removed. The total number of parameters introduced by the IB components therefore is 619,049, which is 0.564% of the total number of model parameters. Therefore, it is not an expensive cost to introduce the IB regularization. Notably, a larger model is not necessarily a better model, because the better-performing E-MIB has fewer parameters than the L-MIB. For comparison, we also present the number of parameters of the state-of-the-art model MAG-BERT [35]. It can be seen that C-MIB has fewer parameters than MAG-BERT but still reaches a better result, which further demonstrates the effectiveness of C-MIB.

V. Conclusion

We propose three MIB variants, namely L-MIB, E-MIB, and C-MIB, which focus on different perspectives of multimodal representation learning. MIB aims to learn minimal sufficient unimodal/multimodal representation,

and is compatible with any fusion mechanisms which enables flexibility. Our method achieves state-of-the-art performance on two multimodal tasks. Moreover, the visualization suggests that the IB principle helps find more discriminative multimodal representation. We also use various objectives to replace the minimal information constraint in the MIB, all of which achieve promising result.

References

[1] P. P. Liang, Z. Liu, A. Zadeh, and L. P. Morency, "Multimodal language analysis with recurrent multistage fusion," in EMNLP, 2018, pp. 150–161.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb 2019.

[3] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic- and visual-lstm with channel-aware temporal convolution network," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1424–1437, 2021.

[4] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in EMNLP, 2017, pp. 1114–1125.

[5] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," ArXiv, vol. abs/1706.03762, 2017.

[6] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in ACL, Jul. 2019, pp. 6558–6569.

[7] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in ACL, Jul. 2019, pp. 481–492.

[8] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," arXiv preprint physics/0004057, 2000.

[9] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in International Conference on Learning Representations, 2017.

[10] S. Wilson and C. K. Mohan, "An information bottleneck approach to optimize the dictionary of visual data," IEEE Transactions on Multimedia, vol. 20, no. 1, pp. 96–106, 2018.

[11] S. Mai, H. Hu, and S. Xing, "A unimodal representation learning and recurrent decomposition fusion structure for utterance-level multimodal embedding learning," IEEE Transactions on Multimedia, pp. 1–1, 2021.

[12] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017.

[13] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in ICDM, 2016, pp. 439–448.

[14] L. Pang, S. Zhu, and C. W. Ngo, "Deep multimodal learning for affective analysis and retrieval," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2008–2020, 2015.

[15] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," IEEE Intelligent Systems, vol. 28, no. 3, pp. 46–53, 2013.

[16] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in Signal and Information Processing Association Summit and Conference, 2012, pp. 1–4.

[17] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10–21, 2010.

[18] B. Nojavanasghari, D. Gopinath, J. Koushik, and L. P. Morency, "Deep multimodal fusion for persuasiveness prediction," in ICMI, 2016, pp. 284–288.

[19] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," in ACL short paper, 2018.

[20] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," IEEE Intelligent Systems, vol. 31, no. 6, pp. 82–88, 2016.

[21] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in ACL, 2019, pp. 1569–1576.

[22] Z. Liu, Y. Shen, P. P. Liang, A. Zadeh, and L. P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in ACL, 2018, pp. 2247–2256.

[23] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," IEEE Transactions on Multimedia, vol. 22, no. 1, pp. 122–137, 2020.

[24] H. Pham, P. P. Liang, T. Manzini, L. P. Morency, and P. Barnabăs, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in AAAI, 2019, pp. 6892–6899.

[25] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in AAAI, vol. 34, no. 01, 2020, pp. 164–172.

[26] A. Zadeh, P. P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L. P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in ACL, 2018, pp. 2236–2246.

[27] S. Mai, S. Xing, J. He, Y. Zeng, and H. Hu, "Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion," arXiv preprint arXiv:2011.13572, 2020.

[28] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in NAACL, 2019, pp. 370–379.

[29] A. Shenoy and A. Sardana, "Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation," arXiv preprint arXiv:2002.08267, 2020.

[30] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in AAAI, vol. 33, 2019, pp. 7216–7223.

[31] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L. P. Morency, "Multi-attention recurrent network for human communication comprehension," in AAAI, 2018, pp. 5642–5649.

[32] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," Information Fusion, vol. 65, pp. 58 – 71, 2021.

[33] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," arXiv preprint arXiv:2001.08735, 2020.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in NAACL-HLT, 2019.

[35] W. Rahman, M. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," ACL, vol. 2020, pp. 2359–2369, 2020.

[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in NeurIPS, 2019.

[37] K. Yang, H. Xu, and K. Gao, "Cm-bert: Cross-modal bert for text-audio sentiment analysis," in ACM MM, 2020, pp. 521–528.

[38] X. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow," arXiv preprint arXiv:1810.00821, 2019.

[39] A. Goyal, R. Islam, D. Strouse, Z. Ahmed, M. Botvinick, H. Larochelle, S. Levine, and Y. Bengio, "Infobot: Transfer and exploration via the information bottleneck," arXiv preprint arXiv:1901.10902, 2019.

[40] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich, "Learning efficient multi-agent communication: An information bottleneck approach," in ICML. PMLR, 2020, pp. 9908–9918.

[41] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," arXiv preprint arXiv:2002.07017, 2020.

[42] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural computation, vol. 14, no. 8, pp. 1771–1800, 2002.

[43] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in NeurIPS, 2018, pp. 5580–5590.

[44] C. Lee and 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.

[45] Y. Zhu, J. Xie, and Z. Chen, "Predicting the popularity of micro-videos with multimodal variational encoder-decoder framework," arXiv preprint arXiv:2003.12724, 2020.

[46] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, "Deep multi-view information bottleneck," in Proceedings of the 2019 SIAM International Conference on Data Mining. SIAM, 2019, pp. 37–45.

[47] S. Ma, D. McDuff, and Y. Song, "Unpaired image-to-speech synthesis with multimodal information bottleneck," in CVPR, 2019, pp. 7598–7607.

[48] D. Olson, "From utterance to text: The bias of language in speech and writing," Harvard Educational Review, vol. 47, no. 3, pp. 257–281, 1977.

[49] A. Shapiro, "Monte carlo sampling methods," Handbooks in operations research and management science, vol. 10, pp. 353–425, 2003.

[50] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," IEEE Intelligent Systems, vol. 31, no. 6, pp. 82–88, 11 2016.

[51] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[52] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, "Memory fusion network for multi-view sequential learning," in AAAI, 2018, pp. 5634–5641.

[53] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," ACM MM, 2020.

[54] Z. Yuan, W. Li, H. Xu, and 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. Dai, Z. Liu, T. Yu, and P. Fung, "Modality-transferable emotion embeddings for low-resource multimodal emotion recognition," in AACL-IJCNLP, Dec. 2020, pp. 269–280.

[56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." Journal of machine learning research, vol. 9, no. 11, 2008.

[57] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in ECCV. Springer, 2016, pp. 443–450.