Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations (Appendix)

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

I. Model Architecture

In this section, we introduce the unimodal learning networks, the multimodal fusion network, and the deduction of the MIB objective in detail. The detailed introduction is shown in the following subsections:

A. *Unimodal Learning Network*: F^m

The input to the model is an utterance [1], 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 $\boldsymbol{U_a} \in \mathbb{R}^{T_a \times d_a}$, $\boldsymbol{U_v} \in \mathbb{R}^{T_v \times d_v}$, 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 ($m \in \{a, v, l\}$).

Since Transformer-based [2] structures enable parallel computation in the time dimension and can learn longer temporal dependency in long sequences, we apply Transformer-based [2] architectures to build up the unimodal learning networks. Specifically, for acoustic and visual modalities, we utilize a standard Transformer to learn the high-level unimodal representations. For the multimodal sentiment analysis task, following the state-of-the-art methods [3]–[5], the large-pretrained Transformer-based model, i.e., BERT [6] is applied to extract the high-level language representation. While for multimodal emotion recognition task, we use GloVe word embeddings [7] to extract language embedding following the state-of-the-art methods [8], [9], and then use Transformer to further learn language representation.

Specifically, the procedures of the BERT unimodal learning network are shown as below:

$$\begin{aligned} \hat{\boldsymbol{X}}^l &= \text{BERT}(\boldsymbol{U}^l) \\ \boldsymbol{X}^l &= \text{Conv 1D}\left(\hat{\boldsymbol{X}}^l, K_l\right) \in \mathbb{R}^{T_l \times d} \\ \boldsymbol{x}^l &= \boldsymbol{X}^l_{T_l} \in \mathbb{R}^d \end{aligned} \tag{1}$$

where Conv 1D denotes the temporal convolution operation with K_l being the kernel size, which is used for mapping the output dimensionality of BERT to the shared dimensionality d that is equal for all modalities. Note that $\boldsymbol{x}^l$ is the feature embedding of $\boldsymbol{X}^l$ in the last time step, and we only use the feature embedding of the last time step to conduct fusion and prediction such that our model is suitable for handling the fusion of unimodal sequences of various length. For the Transformer unimodal learning network, the procedures are presented as follows:

$$\begin{aligned} \boldsymbol{X}^m &= \text{Transformer}(\boldsymbol{U}^m) \in \mathbb{R}^{T_m \times d},\ m \in \{a, v, l\} \\ \boldsymbol{x}^m &= \boldsymbol{X}^m_{T_m} \in \mathbb{R}^d \end{aligned} \tag{2}$$

Note that we only use the encoder of the Transformer to build up the unimodal network.

In our L-MIB and C-MIB, the unimodal representations are regularized by the information bottleneck principle such that the noisy information can be filtered out and the distribution gap between different modalities can be reduced for better fusion. The MIB regularization is demonstrated in Section I-C.

B. *Multimodal Fusion Network*: F^f

Our algorithm is independent of the concrete fusion mechanism, and we can inject various fusion methods into our multimodal fusion structure to provide higher expressive power. In this paper, we mainly investigate five fusion methods to verify the effectiveness of our algorithm. The fusion methods are illustrated as follows:

1) **Direct Addition**: The equation is presented as follows:

$$\boldsymbol{x} = \boldsymbol{x}^l + \boldsymbol{x}^a + \boldsymbol{x}^v \tag{3}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the multimodal representation. This method of fusion is not learnable. In our experiment, we show that even with such a simple fusion method, our algorithm can still reach very competitive performance.

2) **Multiplication**: The equation is presented as follows:

$$\boldsymbol{x} = \boldsymbol{x}^l \cdot \boldsymbol{x}^a \cdot \boldsymbol{x}^v \tag{4}$$

Multiplication is another non-parametric fusion method.

3) **Concatenation**: The equation is presented as follows:

$$\boldsymbol{x} = \boldsymbol{x}^l \oplus \boldsymbol{x}^a \oplus \boldsymbol{x}^v \tag{5}$$

We then use a fully-connected network to map the feature dimensionality of $\boldsymbol{x}$ to d. Together with Direct Addition and Multiplication, they serve as the baseline fusion methods throughout the researches of multimodal learning.

4) **Tensor Fusion**: Tensor fusion [10] is a widely-used fusion algorithm that has attracted significant research attention

Haifeng Hu (corresponding author) is with the School of Electronics and Information Technology, Sun Yat-sen University, China.
E-mail: huhaif@mail.sysu.edu.cn
Sijie Mai and Ying Zeng are with the School of Electronics and Information Technology, Sun Yat-sen University, China.
[1]These authors contribute equally.

recently [11]–[13]. By applying outer product to explore the interactions between unimodal representations, the generated multimodal representation has the highest expressive power but meanwhile is high-dimensional and complicated. The equations for tensor fusion are shown below:

$$\boldsymbol{x}^{m'} = [\boldsymbol{x}^{m},\ 1], \quad m \in \{l, v, a\} \tag{6}$$

$$\boldsymbol{x} = \bigotimes_{m} \boldsymbol{x}^{m'}, \quad \boldsymbol{x}^{m'} \in \mathbb{R}^{d+1} \tag{7}$$

where $\bigotimes$ denotes outer product of a set of vectors. We then use a fully-connected network to map the feature dimensionality of $\boldsymbol{x}$ to d. In Eq. 6, each unimodal embedding is padded with *1s* to retain interactions of any subsets of modalities as in [10].

5) **Graph Fusion**: Graph fusion [14] regards each interaction as one node, and conducts message passing between nodes to model unimodal, bimodal, and trimodal dynamics. The final graph representation is obtained by averaging the node embedding. For more details, please refer to the Graph Fusion Network in [14].

In our E-MIB and C-MIB, the fused multimodal representation is regularized by the IB principle so that it can be sufficient to predict the label and meanwhile contain as less noisy information from the three modalities as possible. The MIB regularization is demonstrated in the following section.

C. Deduction of the MIB objective

In this section, we introduce the deduction of the MIB objective in detail (take E-MIB as an example).

*1) **Introduction of Information Bottleneck**:* Firstly, we introduce the general principle of information bottleneck. Information Bottleneck (IB) is based on Mutual information (MI), aiming to maximize the MI between the encoded representation and the labels as well as simultaneously 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:

$$\begin{aligned} I(\boldsymbol{x}; y) &= I(y; \boldsymbol{x}) \\ &= KL\Big(p(\boldsymbol{x}, y) || p(\boldsymbol{x})p(y)\Big) \\ &= \int d\boldsymbol{x} dy p(\boldsymbol{x}, y) \log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)} \\ &= \mathbb{E}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y)}\left[\log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)}\right] \end{aligned} \tag{8}$$

A common objective function for IB can be defined as:

$$L_{IB} = I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}; \boldsymbol{z}) \tag{9}$$

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

Note that we assume that the joint distribution $p(\boldsymbol{x}, y, \boldsymbol{z})$ can be factorized as follows:

$$p(\boldsymbol{x}, y, \boldsymbol{z}) = p(\boldsymbol{z} \mid \boldsymbol{x}, y)p(y \mid \boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{z} \mid \boldsymbol{x})p(y \mid \boldsymbol{x})p(\boldsymbol{x}) \tag{10}$$

Here we assume that $p(\boldsymbol{z} \mid \boldsymbol{x}, y)$ is equal to $p(\boldsymbol{z} \mid \boldsymbol{x})$, which is reasonable because the encoded representation $\boldsymbol{z}$ only depends on the input $\boldsymbol{x}$ and is independent of the target y.

*2) **Deduction of MIB**:* The objective function of E-MIB is defined as:

$$\begin{aligned} L_{E-MIB} &= I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}^a, \boldsymbol{x}^v, \boldsymbol{x}^l; \boldsymbol{z}) \\ &= I(y; \boldsymbol{z}) - \beta I(\boldsymbol{x}; \boldsymbol{z}) \\ &\text{s.t. } \boldsymbol{x} = F^f(\boldsymbol{x}^a, \boldsymbol{x}^v, \boldsymbol{x}^l;\ \theta_f) \end{aligned} \tag{11}$$

To optimize the objective function of the E-MIB, we adopt the solution introduced in the Variational Information Bottlenecks (VIB) [15]. Firstly we introduce how we optimize the first term of the E-MIB, i.e., $I(y; \boldsymbol{z})$. According to Eq. 8, we can write out $I(y; \boldsymbol{z})$ as

$$\begin{aligned} I(y; \boldsymbol{z}) &= \int dy d\boldsymbol{z} p(y, \boldsymbol{z}) \log \frac{p(y, \boldsymbol{z})}{p(y)p(\boldsymbol{z})} \\ &= \int dy d\boldsymbol{z} p(y, \boldsymbol{z}) \log \frac{p(y \mid \boldsymbol{z})}{p(y)} \\ &= \int dy d\boldsymbol{z} p(y, \boldsymbol{z}) \log p(y \mid \boldsymbol{z}) - \int dy p(y) \log p(y) \end{aligned} \tag{12}$$

According to Eq. 10, $p(y \mid \boldsymbol{z})$ can be formulated as:

$$\begin{aligned} p(y \mid \boldsymbol{z}) &= \frac{p(y, \boldsymbol{z})}{p(\boldsymbol{z})} = \int d\boldsymbol{x} \frac{p(\boldsymbol{x}, y, \boldsymbol{z})}{p(\boldsymbol{z})} \\ &= \int d\boldsymbol{x} \frac{p(\boldsymbol{z} \mid \boldsymbol{x})p(y \mid \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{z})} \end{aligned} \tag{13}$$

However, the above objective is intractable. Therefore, we define $q(y \mid \boldsymbol{z})$ to be a variational approximation to $p(y \mid \boldsymbol{z})$, which is often assumed as a Gaussian distribution. By applying the property that the KL-divergence of two distributions is greater than or equal to zero, we can obtain a lower bound on $I(y; \boldsymbol{z})$:

$$\begin{aligned} &KL\Big(p(y \mid \boldsymbol{z}) || q(y \mid \boldsymbol{z})\Big) \geq 0 \Longrightarrow \\ &\int dy p(y \mid \boldsymbol{z}) \log p(y \mid \boldsymbol{z}) \geq \int dy p(y \mid \boldsymbol{z}) \log q(y \mid \boldsymbol{z}) \end{aligned} \tag{14}$$

Combining Eq. 12 and Eq. 14, we have the following inequality:

$$\begin{aligned} I(y; \boldsymbol{z}) &\geq \int dy d\boldsymbol{z} p(y, \boldsymbol{z}) \log q(y \mid \boldsymbol{z}) - \int dy p(y) \log p(y) \\ &= \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{z} \mid \boldsymbol{x})p(y \mid \boldsymbol{x})p(\boldsymbol{x}) \log q(y \mid \boldsymbol{z}) \end{aligned} \tag{15}$$

where 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 $I(y; \boldsymbol{z})$.

To optimize the second term of the E-MIB (i.e., the minimal information constraint), according to Eq. 8, we can write out the minimal information constraint as:

$$\begin{aligned} I(\boldsymbol{x}; \boldsymbol{z}) &= \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x}, \boldsymbol{z}) \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{x})p(\boldsymbol{z})} \\ &= \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x}, \boldsymbol{z}) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{p(\boldsymbol{z})} \end{aligned} \tag{16}$$

We assume $q(\boldsymbol{z})$ to be a variational approximation to the marginal distribution $p(\boldsymbol{z})$ which is often fixed to a standard normal Gaussian distribution. Similarly, by applying the property that the KL-divergence of two distributions is greater than or equal to zero, we can obtain an upper bound of $I(\boldsymbol{x};\boldsymbol{z})$, as shown below:

$$\begin{aligned} &KL\Big(p(\boldsymbol{z})||q(\boldsymbol{z})\Big) \geq 0 \Longrightarrow \\ &\int d\boldsymbol{z} p(\boldsymbol{z}) \log p(\boldsymbol{z}) \geq \int d\boldsymbol{z} p(\boldsymbol{z}) \log q(\boldsymbol{z}) \end{aligned} \tag{17}$$

Combining Eq. 16 and Eq. 17, we have the following inequality:

$$\begin{aligned} I(\boldsymbol{x};\boldsymbol{z}) &= \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x},\boldsymbol{z}) \log p(\boldsymbol{z} \mid \boldsymbol{x}) - \int d\boldsymbol{z} p(\boldsymbol{z}) \log p(\boldsymbol{z}) \\ &\leq \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x},\boldsymbol{z}) \log p(\boldsymbol{z} \mid \boldsymbol{x}) - \int d\boldsymbol{z} p(\boldsymbol{z}) \log q(\boldsymbol{z}) \\ &= \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x}) p(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \end{aligned} \tag{18}$$

Combining the above two constraints and applying Eq. 10, we can obtain a lower bound of the objective function of E-MIB, which is presented as follows:

$$\begin{aligned} L_{E-MIB} &= I(y;\boldsymbol{z}) - \beta \cdot I(\boldsymbol{x};\boldsymbol{z}) \\ &\geq \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{z} \mid \boldsymbol{x}) p(y \mid \boldsymbol{x}) p(\boldsymbol{x}) \log q(y \mid \boldsymbol{z}) \\ &\quad - \beta \int d\boldsymbol{x} d\boldsymbol{z} p(\boldsymbol{x}) p(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \\ &= \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{z} \mid \boldsymbol{x}) p(\boldsymbol{x}, y) \log q(y \mid \boldsymbol{z}) - \\ &\quad \beta \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{z} \mid x) p(\boldsymbol{x}, y) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \\ &= \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{z}, \boldsymbol{x}, y) \log q(y \mid \boldsymbol{z}) - \\ &\quad \beta \int d\boldsymbol{x} dy d\boldsymbol{z} p(\boldsymbol{x}, y) p(\boldsymbol{z} \mid \boldsymbol{x}) \log \frac{p(\boldsymbol{z} \mid \boldsymbol{x})}{q(\boldsymbol{z})} \\ &= \mathbb{E}_{(\boldsymbol{x},y)\sim p(\boldsymbol{x},y), \boldsymbol{z}\sim p(\boldsymbol{z}|\boldsymbol{x})} \Bigg[\log q(y \mid \boldsymbol{z}) - \\ &\quad \beta \cdot KL\Big(p(\boldsymbol{z} \mid \boldsymbol{x})||q(\boldsymbol{z})\Big) \Bigg] \\ &= J_{E-MIB} \end{aligned} \tag{19}$$

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.

The deduction of L-MIB and C-MIB are similar to that of E-MIB, and is omitted here.

II. EXPERIMENTS

In this section, we present the detailed experimental setting, evaluation protocol, and the introduction of baselines.

A. Evaluation Protocol

For multimodal sentiment analysis, we adopt the following metrics to evaluate the performance of each model: 1) Acc7: 7-class accuracy, sentiment score classification (from -3 to +3); 2) Acc2: binary sentiment score classification, positive or negative; 3) F1: F1 score of the 2-way sentiment classification; 4) MAE: mean absolute error between the prediction and the true sentiment, and 5) Corr: the correlation between the model's prediction and that of humans. For the multimodal emotion recognition task on CMU-MOSEI [16], we follow prior works [8], [17] to report the weighted accuracy and auc score of each emotion. For IEMOCAP dataset [18], we report the accuracy and F1 score of each emotion.

B. Baselines

1) **Early Fusion LSTM** (**EF-LSTM**), which concatenates the features of different modalities at word-level, and then sends the concatenated features to an LSTM layer followed by a classifier to make prediction. 2) **Late Fusion LSTM** (**LF-LSTM**) uses an LSTM network for each modality to extract unimodal features and infer decision, and then combine the unimodal decisions by voting mechanism. 3) **Recurrent Attended Variation Embedding Network** (**RAVEN**) [19], which models interactions by shifting language representations based on the features of the acoustic and visual modalities. 4) **Memory Fusion Network** (**MFN**) [20], which proposes delta-attention module and multi-view gated memory network to discover inter-modal interactions. 5) **Multimodal Transformer** (**MULT**) [21], which learns multimodal representation by translating source modality into target modality using cross-modal Transformer [2]. 6) **Interpretable Modality Fusion** (**IMR**) [22], which improves the interpretable ability of MULT by introducing the multimodal routing mechanism. 7) **Tensor Fusion Network** (**TFN**) [10], which utilizes outer product from unimodal embeddings to jointly learn unimodal, bimodal and trimodal interactions. 8) **Low-rank Modality Fusion** (**LMF**) [12], which leverages low-rank weight tensors to reduce the complexity of TFN. 9) **Quantum-inspired Multimodal Fusion** (**QMF**) [23], which addresses the interpretable ability of multimodal fusion by quantum theory. 10) **Multimodal Adaption Gate BERT** (**MAG-BERT**) [4], which proposes a module called Multimodal Adaptation Gate that enables BERT and XLNet to accept multimodal data during fine-tuning. 11) **Transformer-based Feature Reconstruction Network** (**TFR-Net**) [24], which proposes feature reconstruction network to improve the robustness of multimodal network for the random missing in modality sequences. 12) **Modality-Invariant and -Specific Representation** (**MISA**) [3], which learns unimodal representation by two distinct subspaces mapped for each modality. MISA is currently the state-of-the-art algorithm on CMU-MOSEI dataset (multimodal sentiment analysis task). 13) **Modality-Transferable Emotion Embedding** (**MTEE**) [8] is a multimodal emotion recognition baseline that utilizes the relationship between different emotion categories to improve the performance, which also has a good performance for low-resource or unseen emotions. 14) **Multi-Task Learning** (**MTL**) [17] uses a context-level inter-modal attention framework for simultaneously predicting the sentiment and expressed emotions of an utterance. 15) **Hierarchical feature fusion network** (**HFFN**) [11] applies a

hierarchical 'Divide, Conquer, and Combine' fusion strategy to fuse the three modalities for multimodal learning. 16) **Graph Memory Fusion Network** (**Graph-MFN**) [16] extends MFN [20] by a dynamic graph fusion to fuse the memory of the unimodal LSTMs. 17) **Temporal Convolutional Multimodal LSTM** (**TCM-LSTM**) [9] uses temporal convolutional network to extract the high-level unimodal representations and designs visual-/acoustic-LSTMs for multimodal fusion.

C. Experimental Details

1) Baseline Evaluation: For each baseline of the multimodal sentiment analysis task, following Gkoumas *et al.* [25], we first perform fifty-times random grid search on the hyper-parameters to train the model, and save the hyper-parameter setting that reaches the best performance. After the search of hyper-parameters, we train each model again with the best hyper-parameter setting for five times, and the final results are obtained by calculating the mean results of the five-time running. Notably, for the reason that the codes of QMF [23] and MISA [3] are unavailable at the time of submission, we directly present the results in their original papers. For the baselines of the multimodal emotion recognition task, we borrow the results from [9] and [8].

2) Feature Pre-extraction: For feature pre-extraction, Facet [1] is applied for the visual modality to extract visual features that are composed of facial action units, facial landmarks, head pose, etc. These visual features are extracted from the utterance at the frequency of 30Hz to constitute a sequence of facial representations over time. COVAREP [26] is used for extracting features of the acoustic modality, where the acoustic features include 12 Mel-frequency cepstral coefficients, pitch tracking, speech polarity, glottal closure instants, spectral envelope, etc. These acoustic features are extracted from the full audio clip of the utterance at 100Hz to form a sequence that represents variations in the tone of voice across the utterance. P2FA [27] is used for word-level alignment between the modalities such that the features of the three modalities are aligned at the time dimension. The pre-extracted features are then sent into the unimodal learning networks introduced above to extract high-level unimodal representations. For the multimodal sentiment analysis task, following the state-of-the-art methods [3]–[5], BERT [6] is used to extract the high-level language representation. Following the state-of-the-art methods in multimodal emotion recognition [8], [9], [21], GloVe word embeddings [7] are used to extract the features of the transcripts in the videos. For multimodal emotion recognition task, the input feature dimensionality of language, acoustic, and visual modality is 300, 74, and 35, respectively. For CMU-MOSEI dataset (multimodal sentiment analysis task), the dimensionality of the language feature is 768. For CMU-MOSI, the input dimensionality of language, acoustic, and visual modality is 768, 74, and 47, respectively. The sequence length is set to 50 for CMU-MOSI and CMU-MOSEI, and 20 for IEMOCAP.

3) Hyperparameter Setting: We develop our model using the PyTorch framework on RTX2080Ti with CUDA 10.1 and torch 1.1.0 as the framework. Our proposed model is trained using the optimizer Adam [28]. The learning rate is set to 1e-5 for multimodal sentiment analysis task and 1e-3 for multimodal emotion recognition task. The defaulted fusion method is set to Concatenation for all the MIB variants. The β is set to 1e-3 for all the MIB variants. The encoded feature dimensionality d is set to 50.

[1] iMotions 2017. https://imotions.com/

REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

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

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

[7] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Advances in Neural Information Processing Systems*, 2019, pp. 12 113–12 122.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.
[23] 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.
[24] 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.
[25] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. wei Song, "What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis," *Information Fusion*, vol. 66, pp. 184–197, 2021.
[26] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep: A collaborative voice analysis repository for speech technologies," in *ICASSP*, 2014, pp. 960–964.
[27] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Acoustical Society of America Journal*, vol. 123, p. 3878, 2008.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.