MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis

Devamanyu Hazarika
School of Computing
National University of Singapore
hazarika@comp.nus.edu.sg

Roger Zimmermann
School of Computing
National University of Singapore
rogerz@comp.nus.edu.sg

Soujanya Poria
ISTD, Singapore University of Technology and Design
sporia@sutd.edu.sg

ABSTRACT

Multimodal Sentiment Analysis is an active area of research that leverages multimodal signals for affective understanding of user-generated videos. The predominant approach, addressing this task, has been to develop sophisticated fusion techniques. However, the heterogeneous nature of the signals creates distributional modality gaps that pose significant challenges. In this paper, we aim to learn effective modality representations to aid the process of fusion. We propose a novel framework, MISA, which projects each modality to two distinct subspaces. The first subspace is modality-invariant, where the representations across modalities learn their commonalities and reduce the modality gap. The second subspace is modality-specific, which is private to each modality and captures their characteristic features. These representations provide a holistic view of the multimodal data, which is used for fusion that leads to task predictions. Our experiments on popular sentiment analysis benchmarks, MOSI and MOSEI, demonstrate significant gains over state-of-the-art models. We also consider the task of Multimodal Humor Detection and experiment on the recently proposed UR_FUNNY dataset. Here too, our model fares better than strong baselines, establishing MISA as a useful multimodal framework.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Multimedia information systems*; **Sentiment analysis**.

KEYWORDS

multimodal sentiment analysis; multimodal representation learning

ACM Reference Format:
Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3394171.3413678

1 INTRODUCTION

With the abundance of user-generated online content, such as videos, *Multimodal Sentiment Analysis* (MSA) of human spoken language has become an important area of research [33, 45]. Unlike traditional affect learning tasks performed on isolated modalities (such as text, speech), multimodal learning leverages multiple sources of information comprising language (text/transcripts/ASR), audio/acoustic, and visual modalities. Most of the approaches in MSA are centered around developing sophisticated fusion mechanisms, which span from attention-based models to tensor-based fusion [41]. Despite the advances, these fusion techniques are often challenged by the modality gaps persisting between heterogeneous modalities. Additionally, we want to fuse complementary information to minimize redundancy and incorporate a diverse set of information. One way to aid multimodal fusion is to first learn latent modality representations that capture these desirable properties. To this end, we propose MISA, a novel multimodal framework that learns factorized subspaces for each modality and provides better representations as input to fusion.

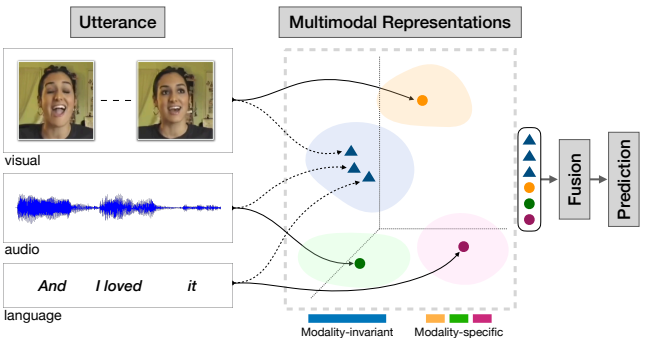


Figure 1: Learning multimodal representations through modality-invariant and -specific subspaces. These features are later utilized for fusion and subsequent prediction of affect in the video.

Motivated by recent advances in domain adaptation [5], MISA learns two distinct utterance representations for each modality. The first representation is *modality-invariant* and aimed towards reducing modality gaps. Here, all the modalities for an utterance are mapped to a shared subspace with distributional alignment. Though multimodal signals come from different sources, they share common motives and goals of the speaker, which is responsible for the overall affective state of the utterance. The invariant mappings help capture these underlying commonalities and correlated features as aligned projections on the shared subspace. Most of the prior works do not utilize such alignment before fusion, which puts an extra burden on their fusion to bridge modality gaps and learn common features.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3413678

In addition to the invariant subspace, MISA also learns *modality-specific* features that are private to each modality. For any utterance, each modality holds distinctive characteristics that include speaker-sensitive stylistic information. Such idiosyncratic details are often uncorrelated to other modalities and are categorized as noise. Nevertheless, they could be useful in predicting the affective state – for example, a speaker's tendency to be sarcastic or peculiar expressions biased towards an affective polarity. Learning such modality-specific features, thus, complements the common latent features captured in the invariant space and provides a comprehensive multimodal representation of the utterance. We propose to use this full set of representations for fusion (see Fig. 1).

To learn these subspaces, we incorporate a combination of losses that include distributional similarity loss (for invariant features), orthogonal loss (for specific features), reconstruction loss (for representativeness of the modality features), and the task prediction loss. We evaluate our hypothesis on two popular benchmark datasets of MSA – MOSI and MOSEI. We also check the adaptability of our model to another similar task – *Multimodal Humor Detection* (MHD), where we evaluate the recently proposed UR_FUNNY dataset. In all three cases, we observe strong gains that surpass state-of-the-art models, highlighting the efficacy of MISA.

The novel contributions of this paper can be summarized as:

- We propose *MISA* – a simple and flexible multimodal learning framework that emphasizes on multimodal representation learning as a pre-cursor to multimodal fusion. MISA learns modality-invariant and modality-specific representations to give a comprehensive and disentangled view of the multimodal data, thus aiding fusion for predicting affective states.
- Experiments on MSA and MHD tasks demonstrate the power of MISA where the learned representations help a simple fusion strategy surpass complex state-of-the-art models.

2 RELATED WORKS

2.1 Multimodal Sentiment Analysis.

The literature in MSA can be broadly classified into: (*i*) *Utterance-level* (*ii*) *Inter-utterance contextual* models. While utterance-level algorithms consider a target utterance in isolation, contextual algorithms utilize neighboring utterances from the overall video.

Utterance-level. Proposed works in this category have primarily focused on learning cross-modal dynamics using sophisticated fusion mechanisms. These works include variety of methods, such as, multiple kernel learning [42], and tensor-based fusion (including its low-rank variants) [15, 21, 26, 29, 31, 58]. While these works perform fusion over representations of utterances, another line of work takes a fine-grained view to perform fusion at the word level. Approaches include multimodal-aware word embeddings [56], recurrent multi-stage fusion [24], graph-based fusion [30, 60], recurrent networks (RNNs), attention-models, memory mechanisms, and transformer-based models [8, 46, 46, 52, 56, 59–61].

Inter-utterance context. These models utilize the context from surrounding utterances of the target utterance. Designed as hierarchical networks, they model individual utterances at the lower level and inter-utterance sequential information in the second level. Poria et al. proposed one of the first models, *bc-LSTM*, which utilized this design along with bi-directional LSTMs for the inter-utterance representation learning, framing the overall problem as a structured prediction (sequence tagging) task [44]. Later works involved either improving fusion using attention [7, 17, 43], hierarchical fusion [32], or developing better contextual modeling [2, 6, 7, 16].

Our work is fundamentally different from these available works. We do not use contextual information and neither focus on complex fusion mechanisms. Instead, we stress the importance of representation learning before fusion. Nevertheless, MISA is flexible to incorporate these above-mentioned components, if required.

2.2 Multimodal Representation Learning.

Common subspace representations. Works that attempt to learn cross-modal common subspaces can be broadly categorized into: (*i*) *Translation*-based models which translates one modality to another using methods such as sequence-to-sequence [40], cyclic translations [39], and adversarial auto-encoders [30]; (*ii*) Correlation-based models [50] that learn cross-modal correlations using Canonical Correlation Analysis [3]; (*iii*) Learning a new shared subspace where all the modalities are simultaneously mapped, using techniques such as adversarial learning [35, 37]. Similar to the third category, we also learn common modality-invariant subspaces. However, we do not use adversarial discriminators to learn shared mappings. Moreover, we incorporate orthogonal modality-specific representations – a trait less explored in multimodal learning tasks.

Factorized representations. Within the regime of subspace learning, we turn our focus to factorized representations. While one line of work attempts to learn generative-discriminative factors of the multimodal data [51], our focus is to learn modality-invariant and -specific representations. To achieve this, we take motivation from related literature on shared-private representations.

The origins of shared-private [5] learning can be found in multi-view component analysis [48]. These early works designed latent variable models (LVMs) with separate shared and private latent variables [9]. Wang et al. [55] revisited this framework by proposing a probabilistic CCA – deep variational CCA. Unlike these models, our proposal involves a discriminative deep neural architecture that obviates the need for approximate inference.

Our framework is closely related to the *Domain Separation Network* (DSN) [5], which proposed the shared-private model for domain adaptation. DSN has been influential in the development of similar models in areas such as multi-task text classification [25]. Although we derive inspiration from DSN, MISA contains critical distinctions: (*i*) DSN learns factorized representations across instances, whereas MISA learns the representations for modalities within instances (utterances); (*ii*) Unlike DSN, we use a more-advanced distribution similarity metric – CMD (see Section 3.5) over adversarial training or MMD; (*iii*) We incorporate additional orthogonal losses across modality-specific (private) representations (see Section 3.5.2); (*iv*) Finally, while DSN uses only shared representations for task predictions, MISA incorporates both invariant and specific representations for fusion followed by task prediction. We posit that availing both the modality representations helps aid fusion by providing a holistic view of the multimodal data.

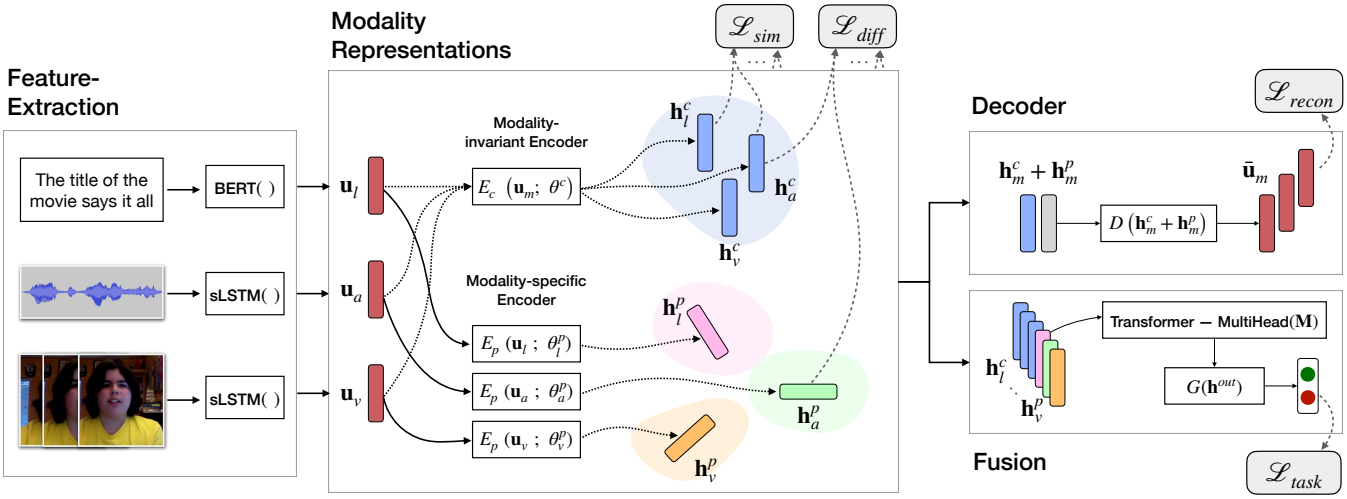


Figure 2: MISA takes the utterance-level representations and projects each modality to two subspaces: modality-invariant and -specific. Later, these hidden representations are used to reconstruct each input and also used for fusion to make the task predictions.

3 APPROACH

3.1 Task Setup

Our goal is to detect sentiments in videos by leveraging multimodal signals. Each video in the data is segmented into its constituent utterances [1], where each utterance — a smaller video by itself — is considered as an input to the model. For an utterance U, the input comprises of three sequences of low-level features from language (l), visual (v) and acoustic (a) modalities. These are represented as $\mathbf{U}_l \in \mathbb{R}^{T_l \times d_l}$, $\mathbf{U}_v \in \mathbb{R}^{T_v \times d_v}$, and $\mathbf{U}_a \in \mathbb{R}^{T_a \times d_a}$ respectively. Here T_m denotes the length of the utterance, such as number of tokens (T_l), for modality m and d_m denotes the respective feature dimensions. The details of these features are discussed in Section 4.3.

Given these sequences $\mathbf{U}_{m \in \{l,v,a\}}$, the primary task is to predict the affective orientation of utterance U from either a predefined set of C categories $y \in \mathbb{R}^C$ or as a continuous intensity variable $y \in \mathbb{R}$.

3.2 MISA

The functioning of MISA can be segmented into two main stages: Modality Representation Learning (Section 3.3) and Modality Fusion (Section 3.4). The full framework is illustrated in Fig. 2.

3.3 Modality Representation Learning

Utterance-level Representations. Firstly, for each modality $m \in \{l, v, a\}$, we map its utterance sequence $\mathbf{U}_m \in \mathbb{R}^{T_m \times d_m}$ to a fixed-sized vector $\mathbf{u}_m \in \mathbb{R}^{d_h}$. We use a stacked bi-directional Long Short-Term Memory (LSTM) [20] whose end-state hidden representations coupled with a fully connected dense layer gives $\mathbf{u}_m$:

$$\mathbf{u}_m = \text{sLSTM}\left(\mathbf{U}_m;\ \theta_m^{lstm}\right) \tag{1}$$

Modality-Invariant and -Specific Representations. We now project each of the utterance vector $\mathbf{u}_m$ to two distinct representations. First is the modality-invariant component that learns a shared representation in a common subspace with distributional similarity constraints [18]. This constraint aids in minimizing the heterogeneity gap – a desirable property for multimodal fusion. Second is the modality-specific component that captures the unique characteristics of that modality. Through this paper, we argue that the presence of both modality-invariant and -specific representations provides a holistic view that is required for effective fusion. Learning these representations is the primary goal of our work.

Given the utterance vector $\mathbf{u}_m$ for modality m, we learn the hidden modality-invariant ($\mathbf{h}_m^c \in \mathbb{R}^{d_h}$) and modality-specific ($\mathbf{h}_m^p \in \mathbb{R}^{d_h}$) representations using the encoding functions:

$$\mathbf{h}_m^c = E_c\left(\mathbf{u}_m; \theta^c\right), \quad \mathbf{h}_m^p = E_p\left(\mathbf{u}_m; \theta_m^p\right) \tag{2}$$

To generate the six hidden vectors $\mathbf{h}_{l/v/a}^{p/c}$ (two per modality), we use simple feed-forward neural layers; E_c shares the parameters θ^c across all three modalities, whereas E_p assigns separate parameters θ_m^p for each modality.

3.4 Modality Fusion

After projecting the modalities into their respective representations, we fuse them into a joint vector for downstream predictions. We design a simple fusion mechanism that first performs a self-attention — based on the Transformer [54] — followed by a concatenation of all the six transformed modality vectors.

Definition Transformer. *The Transformer leverages an attention module that is defined as a scaled dot-product function:*

$$\textit{Attention}(Q, K, V) = \textit{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_h}}\right)\mathbf{V} \tag{3}$$

Where, $\mathbf{Q}$, $\mathbf{K}$, *and* $\mathbf{V}$ *are the query, key, and value matrices. The Transformer computes multiple such parallel attentions, where each attention output is called a head. The* i^{th} *head is computed as:*

$$head_i = \textit{Attention}(\mathbf{Q}W_i^q, \mathbf{K}W_i^k, \mathbf{V}W_i^v) \tag{4}$$

[1] An utterance is a unit of speech bounded by breaths or pauses [34].

$W_i^{q/k/v} \in \mathbb{R}^{d_h \times d_h}$ *are head-specific parameters to linearly project the matrices into local spaces.*

Fusion Procedure. First we stack the six modality representations (from Eq. (2)) into a matrix $\mathbf{M} = [\mathbf{h}_l^c, \mathbf{h}_v^c, \mathbf{h}_a^c, \mathbf{h}_l^p, \mathbf{h}_v^p, \mathbf{h}_a^p] \in \mathbb{R}^{6 \times d_h}$. Then, we perform a multi-headed self-attention on these representations to make each vector *aware* of the fellow cross-modal (and cross subspace) representations. Doing this allows each representation to induce potential information from fellow representations that are synergistic towards the overall affective orientation. Such cross-modality matching has been highly prominent in recent cross-modal learning approaches [22, 23, 27, 49, 57].

For self-attention, we set $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{M} \in \mathbb{R}^{6 \times d_h}$. The Transformer generates a new matrix $\bar{\mathbf{M}} = [\bar{\mathbf{h}}_l^c, \bar{\mathbf{h}}_v^c, \bar{\mathbf{h}}_a^c, \bar{\mathbf{h}}_l^p, \bar{\mathbf{h}}_v^p, \bar{\mathbf{h}}_a^p]$ as:

$$\bar{\mathbf{M}} = \text{MultiHead}(\mathbf{M}; \theta^{att}) = (\text{head}_1 \oplus \cdots \oplus \text{head}_n) W^o \tag{5}$$

where, each head_i here is calculated based on Eq. (4); $\oplus$ represents concatenation; and $\theta^{att} = \{W^q, W^k, W^v, W^o\}$.

Prediction/Inference. Finally, we take the Transformer output and construct a joint-vector using concatenation, $\mathbf{h}^{out} = [\bar{\mathbf{h}}_l^c \oplus \cdots \oplus \bar{\mathbf{h}}_a^p] \in \mathbb{R}^{6d_h}$. The task predictions are then generated by the function $\hat{\mathbf{y}} = G(\mathbf{h}^{out}; \theta^{out})$.

We provide the network topology of the functions $sLSTM()$, $E_c()$, $E_p()$, $G()$ and $D()$ (explained later) in the appendix.

3.5 Learning

The overall learning of the model is performed by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \, \mathcal{L}_{\text{sim}} + \beta \, \mathcal{L}_{\text{diff}} + \gamma \, \mathcal{L}_{\text{recon}} \tag{6}$$

Here, α, β, γ are the interaction weights that determine the contribution of each regularization component to the overall loss $\mathcal{L}$. Each of these component losses are responsible for achieving the desired subspace properties. We discuss them next.

3.5.1 $\mathcal{L}_{sim}$ – ***Similarity Loss.*** Minimizing the *similarity loss* reduces the discrepancy between the shared representations of each modality. This helps the common cross-modal features to be aligned together in the shared subspace. Amongst many choices, we use the *Central Moment Discrepancy* (CMD) [63] metric for this purpose. CMD is a state-of-the-art distance metric that measures the discrepancy between the distribution of two representations by matching their order-wise moment differences. Intuitively, CMD distance decreases as two distributions become more similar.

Definition CMD. *Let X and Y be bounded random samples with respective probability distributions p and q on the interval $[a, b]^N$. The central moment discrepancy regularizer CMD_K is defined as an empirical estimate of the CMD metric, by*

$$CMD_K(X, Y) = \frac{1}{|b-a|} \|\mathbf{E}(X) - \mathbf{E}(Y)\|_2 + \sum_{k=2}^{K} \frac{1}{|b-a|^k} \|C_k(X) - C_k(Y)\|_2 \tag{7}$$

where, $\mathbf{E}(X) = \frac{1}{|X|} \sum_{x \in X} x$ is the empirical expectation vector of sample X and $C_k(X) = \mathbf{E}\left((x - \mathbf{E}(X))^k\right)$ is the vector of all k^{th} order sample central moments of the coordinates of X.

In our case, we calculate the CMD loss between the invariant representations of each pair of modalities:

$$\mathcal{L}_{\text{sim}} = \frac{1}{3} \sum_{\substack{(m_1, m_2) \in \\ \{(l,a),(l,v),\\(a,v)\}}} CMD_K(\mathbf{h}_{m_1}^c, \mathbf{h}_{m_2}^c) \tag{8}$$

Here, we make two important observations: (i) We choose CMD over KL-divergence or MMD because CMD is a popular metric [36] and performs explicit matching of higher-order moments without expensive distance and kernel matrix computations. (ii) *Adversarial loss* is another choice for similarity training, where a discriminator and the shared encoder engage in a minimax game. However, we choose CMD owing to its simple formulation. In contrast, adversarial training demands additional parameters for the discriminator along with added complexities, such as oscillations in training [53].

3.5.2 $\mathcal{L}_{diff}$ – ***Difference Loss.*** This loss is to ensure that the modality-invariant and -specific representations capture different aspects of the input. The non-redundancy is achieved by enforcing a soft orthogonality constraint between the two representations [5, 25, 47]. In a training batch of utterances, let $\mathbf{H}_m^c$ and $\mathbf{H}_m^p$ be the matrices [2] whose rows denote the hidden vectors $\mathbf{h}_m^c$ and $\mathbf{h}_m^p$ for modality m of each utterance. Then the orthogonality constraint for this modality vector pair is calculated as:

$$\left\| \mathbf{H}_m^{c^\top} \mathbf{H}_m^p \right\|_F^2 \tag{9}$$

Here, $\| \cdot \|_F^2$ is the squared Frobenius norm. In addition to the constraints between the invariant and specific vectors, we also add orthogonality constraints between the modality-specific vectors. The overall difference loss is then computed as:

$$\mathcal{L}_{\text{diff}} = \sum_{m \in \{l,v,a\}} \left\| \mathbf{H}_m^{c^\top} \mathbf{H}_m^p \right\|_F^2 + \sum_{\substack{(m_1, m_2) \in \\ \{(l,a),(l,v),\\(a,v)\}}} \left\| \mathbf{H}_{m_1}^{p^\top} \mathbf{H}_{m_2}^p \right\|_F^2 \tag{10}$$

3.5.3 $\mathcal{L}_{recon}$ – ***Reconstruction Loss.*** As the *difference loss* is enforced, there remains a risk of learning trivial representations by the modality-specific encoders. Trivial cases can arise if the encoder function approximates an orthogonal but unrepresentative vector of the modality. To avoid this situation, we add a *reconstruction loss* that ensures the hidden representations to capture details of their respective modality. First, we reconstruct the modality vector $\mathbf{u}_m$ by using a decoder function $\hat{\mathbf{u}}_m = D(\mathbf{h}_m^c + \mathbf{h}_m^p; \theta^d)$. The reconstruction loss is then the *mean squared error* loss between $\mathbf{u}_m$ and $\hat{\mathbf{u}}_m$:

$$\mathcal{L}_{\text{recon}} = \frac{1}{3} \left(\sum_{m \in \{l,v,a\}} \frac{\|\mathbf{u}_m - \hat{\mathbf{u}}_m\|_2^2}{d_h} \right) \tag{11}$$

Where, $\| \cdot \|_2^2$ is the squared L_2-norm.

3.5.4 $\mathcal{L}_{task}$ – ***Task Loss.*** The task-specific loss estimates the quality of prediction during training. For classification tasks, we use the standard *cross-entropy loss* whereas for regression tasks, we use

[2] We transform the matrices to have zero mean and unit l_2 norm.

mean squared error loss. For N_b utterances in a batch, these are calculated as:

$$\mathcal{L}_{\text{task}} = -\frac{1}{N_b}\sum_{i=0}^{N_b} \mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i \qquad \text{for classification} \tag{12}$$

$$= \frac{1}{N_b}\sum_{i=0}^{N_b} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 \qquad \text{for regression} \tag{13}$$

4 EXPERIMENTS

4.1 Datasets

We consider benchmark datasets for both the tasks of MSA and MHD. These datasets provide word-aligned multimodal signals (language, visual, and acoustic) for each utterance.

4.1.1 ***CMU-MOSI.*** The CMU-MOSI dataset [62] is a popular benchmark dataset for research in MSA. The dataset is a collection of YouTube monologues, where a speaker expresses their opinions on topics such as movies. With a total of 93 videos, spanning 89 distance speakers, MOSI contains 2198 subjective utterance-video segments. The utterances are manually annotated with a continuous opinion score between $[-3, 3]$, where -3/+3 represents strongly negative/positive sentiments.

4.1.2 ***CMU-MOSEI.*** The CMU-MOSEI dataset [60] is an improvement over MOSI with higher number of utterances, greater variety in samples, speakers, and topics. The dataset contains 23453 annotated video segments (utterances), from 5000 videos, 1000 distinct speakers and 250 different topics.

4.1.3 ***UR_FUNNY.*** For MHD, we consider the recently proposed UR_FUNNY dataset [19]. Similar to sentiments, generating and perceiving humor also occurs through multimodal channels. This dataset, thus provides multimodal utterances that act as punchlines sampled from TED talks. It also provides associated context for each target utterance and ensures diversity in both speakers and topics. Each target utterance is labeled with a binary label for humor/non-humor instance. Dataset split and training details are available in the appendix.

4.2 Evaluation Criteria

Sentiment intensity prediction in both MOSI and MOSEI datasets are regression tasks with *mean absolute error* (MAE) and *Pearson correlation* (Corr) as the metrics. Additionally, the benchmark also involves classification scores that include, *seven-class accuracy (Acc-7)* ranging from -3 to 3, *binary accuracy (Acc-2)* and *F-Score*. For binary accuracy scores, two distinct approaches have been considered in the past. First is *negative*/*non-negative* classification where the labels for non-negatives are based on scores being ≥ 0 [61]. In recent works, binary accuracy is calculated on the more accurate formulation of *negative*/*positive* classes where negative and positive classes are assigned for < 0 and > 0 sentiment scores, respectively [52]. We report results on both these metrics using the segmentation marker $-/-$ where the left-side score is for *neg./non-neg.* while the right-side score is for *neg./pos.* classification. For UR_FUNNY dataset, the task is a standard binary classification with binary accuracy (Acc-2) as the metric for evaluation [19].

4.3 Feature Extraction

For fair comparisons, we utilize the standard low-level features that are provided by the respective benchmarks and utilized by the state-of-the-art methods.

4.3.1 ***Language Features.*** Traditionally, language modality features has been GloVe [38] embeddings for each token in the utterance. However, following recent works [7], including the state-of-the-art ICCN [50], we utilize the pre-trained `BERT` [11] as the feature extractor for textual utterances. Using `BERT` replaces the $sLSTM(\mathbf{U}_l; \theta_l^{lstm})$ in Eq. (1) with $\texttt{BERT}(U_l; \theta^{bert})$. For UR_FUNNY, however, the state of the art is based on GloVe features. Thus, for fair comparison, we provide results using both GloVe and `BERT`.

While GloVe features are 300 dimensional token embeddings, for `BERT`, we utilize the *BERT-base-uncased* pre-trained model. This model comprises of 12 stacked Transformer layers. Aligned with recent works [1], we choose the utterance vector $\mathbf{u}_l$ to be the average representation of the tokens from the final 768 dimensional hidden state. Unfortunately, for our considered UR_FUNNY version, the original transcripts are not available. Instead, only the GloVe embeddings have been provided. To retrieve the raw texts, we choose the token with the least cosine distance from the GloVe vocabulary for each word embedding. A manual check of 100 randomly sampled utterances validated the quality of this process to retrieve legible original transcripts.

4.3.2 ***Visual Features.*** Both MOSI and MOSEI use Facet [3] to extract facial expression features, which include facial action units and face pose based on the Facial Action Coding System (FACS) [14]. This process is repeated for each sampled frame within the utterance video sequence. For UR_FUNNY, OpenFace [4], a facial behavioral analysis tool, is used to extract features related to the facial expressions of the speaker. The final visual feature dimensions, d_v, are 47 for MOSI, 35 for MOSEI, and 75 for UR_FUNNY.

4.3.3 ***Acoustic Features.*** The acoustic features contain various low-level statistical audio functions extracted from COVAREP [10] – an acoustic analysis framework. Some of the features include 12 Mel-frequency ceptral coefficients, pitch, Voiced/Unvoiced segmenting features (VUV) [12], glottal source parameters [13], and other features related to emotions and tone of speech [4]. The feature dimensions, d_a, are 74 for MOSI/MOSEI and 81 for UR_FUNNY.

4.4 Baselines

We perform a comprehensive comparative study against MISA by considering various baselines as detailed below.

4.4.1 ***Previous Models.*** Numerous methods have been proposed for multimodal learning, especially for sentiment analysis and human language tasks in general. As mentioned in Section 2, these works can be broadly categorized into utterance-level and inter-utterance contextual models. Utterance-level baselines include:

[3] https://imotions.com/platform/

[4] Please refer to [19, 60] and their respective SDKs (https://github.com/A2Zadeh/CMU-MultimodalSDK v1.1.1; https://github.com/ROC-HCI/UR-FUNNY/blob/master/UR-FUNNY-V1.md v1) for a full list of the features. Sampling rates for the acoustic and visual signals are summarized in [52]. Following related works, we align all three modalities based on language modality. This standard procedure makes all three temporal sequences within an utterance to be of equal length, i.e. $T_l = T_v = T_a$

- Networks which perform temporal modeling and fusion of utterances: **MFN** [59], **MARN** [61], **MV-LSTM** [46], **RMFN** [24].
- Models which utilize attention and transformer modules to improve token representations using non-verbal signals: **RAVEN** [56], **MulT** [52].
- Graph-based fusion models: **Graph-MFN** [60].
- Utterance-vector fusion approaches that use tensor-based fusion and low-rank variants: **TFN** [58], **LMF** [26], **LMFN** [31], **HFFN** [29].
- Common subspace learning models that use cyclic translations (**MCTN** [39]), adversarial auto-encoders (**ARGF** [30]), and generative-discriminative factorized representations (**MFM** [51]).

Inter-utterance contextual baselines include:

- RNN-based models: **BC-LSTM** [44], with hierarchical fusion – **CH-Fusion** [32].
- Inter-utterance attention and multi-tasking models: **CIA** [6], **CIM-MTL** [2], **DFF-ATMF** [7].

For detailed descriptions of the models, please refer to the appendix.

4.4.2 ***State of the Art.*** For the task of MSA, the Interaction Canonical Correlation Network (**ICCN**) [50] stands as the state-of-the-art (SOTA) model on both MOSI and MOSEI. ICCN first extracts features from audio and video modality and then fuses with text embeddings to get two outer products, text-audio and text-video. Finally, the outer products are fed to a Canonical Correlation Analysis (CCA) network, whose output is used for prediction.

For MHD, The SOTA is Contextual Memory Fusion Network (**C-MFN**) [19], which extends the MFN model by proposing uni- and multimodal context networks that consider preceding utterances and performs fusion using the MFN model as its backbone. Originally, MFN [59] is a multi-view gated memory network that stores intra- and cross-modal utterance interactions in its memories.

5 RESULTS AND ANALYSIS

5.1 Quantitative Results

5.1.1 ***Multimodal Sentiment Analysis.*** The comparative results for MSA are presented in Table 1 (MOSI) and Table 2 (MOSEI). In both the datasets, MISA achieves the best performance and surpasses the baselines — including the state-of-the-art ICCN — across all metrics (regression and classification combined). Within the results, it can be seen that our model, which is an utterance-level model, fares better than the contextual models. This is an encouraging result as we are able to perform better even with lesser information. Our model also surpasses some of the intricate fusion mechanisms, such as TFN and LFN, which justify the importance of learning multimodal representations preceding the fusion stage.

5.1.2 ***Multimodal Humor Detection.*** Similar trends are observed for MHD (see Table 3), with a highly pronounced improvement over the contextual SOTA, C-MFN. This is true even while using GloVe features for language modality. In fact, our GloVe variant is at par to the BERT-based baselines, such as TFN. This indicates that effective modeling of multimodal representations goes a long way. Humor detection is known to be highly sensitive to the idiosyncratic characteristics of different modalities [19]. Such dependencies are well modeled by our representations, which is reflected in the results.

Models	MOSI				
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-Score (↑)	Acc-7 (↑)
BC-LSTM	1.079	0.581	73.9 / -	73.9 / -	28.7
MV-LSTM	1.019	0.601	73.9 / -	74.0 / -	33.2
TFN	0.970	0.633	73.9 / -	73.4 / -	32.1
MARN	0.968	0.625	77.1 / -	77.0 / -	34.7
MFN	0.965	0.632	77.4 / -	77.3 / -	34.1
LMF	0.912	0.668	76.4 / -	75.7 / -	32.8
CH-Fusion	-	-	80.0 / -	-	-
MFM$^{\otimes}$	0.951	0.662	78.1 / -	78.1 / -	36.2
RAVEN$^{\otimes}$	0.915	0.691	78.0 / -	76.6 / -	33.2
RMFN$^{\otimes}$	0.922	0.681	78.4 / -	78.0 / -	38.3
MCTN$^{\otimes}$	0.909	0.676	79.3 / -	79.1 / -	35.6
CIA	0.914	0.689	79.8 / -	- / 79.5	38.9
HFFN$^{\oslash}$	-	-	- / 80.2	- / 80.3	-
LMFN$^{\oslash}$	-	-	- / 80.9	- / 80.9	-
DFF-ATMF (B)	-	-	- / 80.9	- / 81.2	-
ARGF	-	-	- / 81.4	- / 81.5	-
MulT	0.871	0.698	- / 83.0	- / 82.8	40.0
TFN (B)$^{\diamond}$	0.901	0.698	- / 80.8	- / 80.7	34.9
LMF (B)$^{\diamond}$	0.917	0.695	- / 82.5	- / 82.4	33.2
MFM (B)$^{\diamond}$	0.877	0.706	- / 81.7	- / 81.6	35.4
ICCN (B)	0.860	0.710	- / 83.0	- / 83.0	39.0
MISA (B)	**0.783**	**0.761**	**81.8**† / **83.4**†	**81.7** / **83.6**	**42.3**
Δ_{SOTA}	↓ 0.077	↑ 0.051	↑ 2.0 / ↑ 0.4	↑ 2.6 / ↑ 0.6	↑ 3.3

Table 1: Performances of multimodal models in MOSI. NOTE: (B) means the language features are based on BERT; $^{\otimes}$ from [52]; $^{\oslash}$ from [30]; $^{\diamond}$ from [50]. Final row presents our best model per metric. $^{\dagger}p < 0.05$ under McNemar's Test for binary classification. Here, the statistical significance tests are compared with publicly available models of [26, 51, 58].

Models	MOSEI				
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-Score (↑)	Acc-7 (↑)
MFN$^{\otimes}$	-	-	76.0 / -	76.0 / -	-
MV-LSTM$^{\otimes}$	-	-	76.4 / -	76.4 / -	-
Graph-MFN$^{\otimes}$	0.710	0.540	76.9 / -	77.0 / -	45.0
RAVEN	0.614	0.662	79.1 / -	79.5 / -	50.0
MCTN	0.609	0.670	79.8 / -	80.6 / -	49.6
CIA	0.680	0.590	80.4 / -	78.2 / -	50.1
CIM-MTL	-	-	80.5 / -	78.8 / -	-
DFF-ATMF (B)	-	-	- / 77.1	- / 78.3	-
MulT	0.580	0.703	- / 82.5	- / 82.3	51.8
TFN (B)$^{\diamond}$	0.593	0.700	- / 82.5	- / 82.1	50.2
LMF (B)$^{\diamond}$	0.623	0.677	- / 82.0	- / 82.1	48.0
MFM (B)$^{\diamond}$	0.568	0.717	- / 84.4	- / 84.3	51.3
ICCN (B)	0.565	0.713	- / 84.2	- / 84.2	51.6
MISA (B)	**0.555**	**0.756**	**83.6**† / **85.5**†	**83.8** / **85.3**	**52.2**
Δ_{SOTA}	↓ 0.010	↑ 0.043	↑ 3.1 / ↑ 1.3	↑ 5.0 / ↑ 1.1	↑ 0.6

Table 2: Performances of multimodal models in MOSEI. NOTE: (B) means the language features are based on BERT; $^{\otimes}$ from [60]; $^{\diamond}$ from [50]. Final row presents our best model per metric. $^{\dagger}p < 0.05$ under McNemar's Test for binary classification (compared with publicly available models of [26, 51, 58]).

5.1.3 ***BERT vs. GloVe.*** In our experiments, we observe improvements in performance when using BERT over the traditional GloVe-based features for language. This raises the question as to whether our performance improvements are *solely* due to BERT features. To find an answer, we look at the state-of-the-art approach ICCN, which is also based on BERT. Our model comfortably beats ICCN in all metrics, through which we can infer that the improvements in multimodal modeling are a critical factor.

Algorithms	context	target	UR_FUNNY Accuracy-2 (↑)
C-MFN	✓		58.45
C-MFN		✓	64.47
TFN		✓	64.71
LMF		✓	65.16
C-MFN	✓	✓	65.23
LMF (Bert)		✓	67.53
TFN (Bert)		✓	68.57
MISA (GloVe)		✓	68.60
MISA (Bert)		✓	**70.61**†
Δ_{SOTA}			↑ 2.07

Table 3: Performances of multimodal models in UR_FUNNY. $^{\dagger}p < 0.05$ under McNemar's Test for binary classification when compared against [26, 58]. *Context*-based models use additional data that include the utterances preceding the target punchline.

	Model	MOSI		MOSEI		UR_FUNNY
		MAE (↓)	Corr (↑)	MAE (↓)	Corr (↑)	Acc-2 (↑)
1)	MISA	**0.783**	**0.761**	**0.555**	**0.756**	**70.6**
2)	(-) language l	1.450	0.041	0.801	0.090	55.5
3)	(-) visual v	0.798	0.756	0.558	0.753	69.7
4)	(-) audio a	0.849	0.732	0.562	0.753	70.2
5)	(-) $\mathcal{L}_{\text{sim}}$	0.807	0.740	0.566	0.751	69.3
6)	(-) $\mathcal{L}_{\text{diff}}$	0.824	0.749	0.565	0.742	69.3
7)	(-) $\mathcal{L}_{\text{recon}}$	0.794	0.757	0.559	0.754	69.7
8)	*base*	0.810	0.750	0.568	0.752	69.2
9)	*inv*	0.811	0.737	0.561	0.743	68.8
10)	*sFusion*	0.858	0.716	0.563	0.752	70.1
11)	*iFusion*	0.850	0.735	**0.555**	0.750	69.8

Table 4: Ablation Study. Here, (−) represents removal for the mentioned factors. Model 1 represents the best performing model in each dataset; Model 2,3,4 depicts the effect of individual modalities; Model 5,6,7 presents the effect of regularization; Model 8,9,10,11 presents the variants of MISA as defined in Section 5.2.3.

5.2 Ablation Study

5.2.1 ***Role of Modalities.*** In Table 4 (model 2, 3, 4) we remove one modality at a time to observe the effect in performance. Firstly, we see that a multimodal combination provides the best performance, indicating that the model can learn complementary features. Without this case, the tri-modal combination would not fare better than bi-modal variants such as language-visual MISA.

Next, we observe that the performance sharply drops when the language modality is removed. Similar drops are not observed in removing the other two modalities, showing that the text modality has significant dominance over the audio and visual modalities. There could be two reasons for this: 1) The data quality of text modality could be inherently better as they are manual transcriptions. In contrast, audio and visual signals are unfiltered raw signals. 2) BERT is a pre-trained model with better expressive power over the randomly initialized audio and visual feature extractor, giving better utterance-level features. These observations, however, are dataset specific and can not be generalized to any multimodal scenario.

5.2.2 ***Role of Regularization.*** Regularization plays a critical role in achieving the desired representations discussed in Section 3.5.

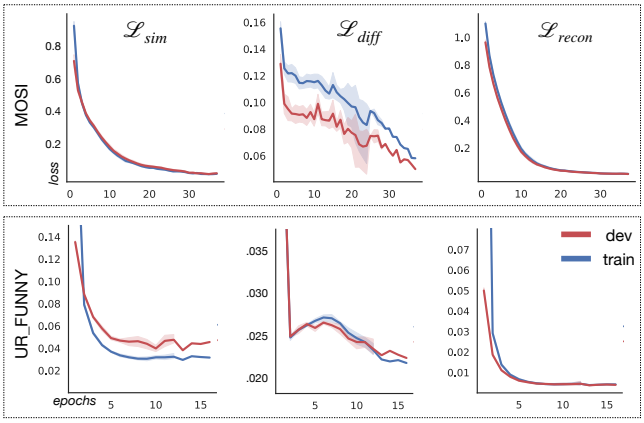


Figure 3: Trends in the regularization losses as training proceeds (values are for five runs across random seeds). Graphs depict losses in both training and validation sets for MOSI and UR_FUNNY. Similar trends are also observed in MOSEI.

In this section, we first observe how well the losses are learned in the model while training and whether the validation sets follow similar trends. Next, we perform qualitative verification by looking at the feature distributions of the learned models. Finally, we look at the importance of each loss by an ablation study.

Regularization Trends. The losses $\{\mathcal{L}_{\text{sim}}, \mathcal{L}_{\text{diff}}, \mathcal{L}_{\text{recon}}\}$ act as measures to quantify how well the model has learnt modality-invariant and -specific representations. We thus trace the losses as training proceeds both in the training and validation sets. As seen in Fig. 3, all three losses demonstrate a decreasing trend with the number of epochs. This shows that the model is indeed learning the representations as per design. Like the training sets, the validation sets also demonstrate similar behavior.

Visualizing Representations. While Fig. 3 shows how regularization losses behave during training, it is also vital to investigate how well these characteristics are generalized. We thus visualize the hidden representations for the samples in the *testing sets*. Fig. 4 presents the illustrations, where it is clearly seen that in the case of no regularization ($\alpha = 0, \beta = 0$), modality-invariance is not learnt. Whereas, when losses are introduced, overlaps amongst the modality-invariant representations are observed. This indicates that MISA is able to perform desired subspace learning, even in the generalized scenario, i.e., in the testing set. We delve further into the utility of these subspaces in Section 5.2.3.

Importance of Regularization. To quantitatively verify the importance of these losses, we take the best models in each dataset and re-train them by ablating one loss at a time. To nullify each loss, we set either $\{\alpha, \beta, \gamma\}$ to 0. Results are observed in Table 4 (model 5,6,7). As seen, the best performance is achieved when all the losses are involved. In a closer look, we can see that the models are particularly sensitive to the *similarity* and *difference* losses that ensures both the modality invariance and specificity. This dependence indicates that having separate subspaces is indeed helpful. For the *reconstruction* loss, we see a lesser dependence on the model. One possibility is that, despite the absence of reconstruction loss,

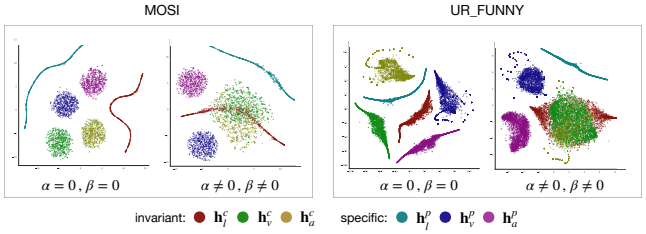


Figure 4: Visualization of the modality-invariant and -specific subspaces in the testing set of MOSI and UR_FUNNY datasets using t-SNE projections [28]. Observations on MOSEI are also similar.

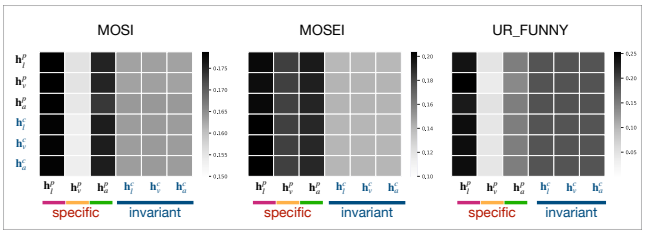


Figure 5: Average self-attention scores from the Transformer-based fusion module. The rows depict the *queries*, columns depict the *keys* (see Section 3.4). Essentially, each column represents the contribution of an input feature vector $\in \{\mathbf{h}_l^c, \mathbf{h}_v^c, \mathbf{h}_a^c, \mathbf{h}_l^p, \mathbf{h}_v^p, \mathbf{h}_a^p\}$ to generate the output feature vectors $[\bar{\mathbf{h}}_l^c, \bar{\mathbf{h}}_v^c, \bar{\mathbf{h}}_a^c, \bar{\mathbf{h}}_l^p, \bar{\mathbf{h}}_v^p, \bar{\mathbf{h}}_a^p]$.

the modality-specific encoders are not resorting to trivial solutions and rather learning informative representations using the task loss. This would not be the case if only the modality-invariant features were used for prediction.

5.2.3 ***Role of subspaces.*** In this section, we look at several variants to our proposed model to investigate alternative hypotheses:

1) *MISA-base* is a baseline version where we do not learn disjoint subspaces. Rather, we utilize three separate encoders for each modality — similar to previous works — and employ fusion on them.
2) *MISA-inv* is a variant where there is no modality-specific representation. In this case, only modality-invariant representations are learnt and subsequently utilized for fusion.
3) The next two variants, *MISA-sFusion* and *MISA-iFusion* are identical to MISA in the representation learning phase. In *MISA-sFusion*, we only use the modality-specific features ($\mathbf{h}^p_{\{l/v/a\}}$) for fusion and prediction. Similarly, *MISA-iFusion* uses only modality-invariant features ($\mathbf{h}^c_{\{l/v/a\}}$) for fusion.

We summarize the results in Table 4 (model 8-11). Overall, we find our final design to be better than the variants. Amongst the variants, we observe that learning only an invariant space might be too restrictive as not all modalities in an utterance share the same polarity stimulus. This is reflected in the results where *MISA-inv* does not fare better than the general *MISA-base* model. Both *MISA-sFusion* and *-iFusion* improve the performances but the best combination is when both representation learning and fusion utilize both the modality subspaces, i.e., the proposed model MISA.

5.2.4 ***Visualizing Attention.*** To analyze the utility of the learned representations, we look at their role in the fusion step. As discussed in Section 3.4, fusion includes a self-attention procedure on the modality representations that enhances each representation $\mathbf{h}^{c/p}_{l/v/a}$ to $\bar{\mathbf{h}}^{c/p}_{l/v/a}$, using a soft-attention combination of all its fellow representations (including itself). Fig. 5 illustrates the average attention distribution of the testing sets. Each row in the figure is a probability distribution for the respective representation (averaged over all the testing samples). Looking at the columns, each column can be seen as the contribution that any vector $\mathbf{h} \in \{\mathbf{h}^{c/p}_{l/v/a}\}$ has to the enhanced representations of all the resulting vectors $\bar{\mathbf{h}}^{c/p}_{l/v/a}$. We observe two important patterns in the figures. First, we notice that the invariant representations influence equally amongst all three modalities. This is true for all the datasets and expected as they are aligned in the shared space. It also establishes that modality gap is reduced amongst the invariant features. Second, we notice a significant contribution from modality-specific representations. Although the average importance of a modality depends on the dataset, language, as also seen in quantitative results, contributes the most while acoustic and visual modalities provide varied levels of influences. Nevertheless, both invariant and specific representations provide information in the fusion, as observed in these influence maps.

6 CONCLUSION

In this paper, we presented MISA – a multimodal affective framework that factorizes modalities into modality-invariant and modality-specific features and then fuses them to predict affective states. Despite comprising simple feed-forward layers, we find MISA to be highly effective and observe significant gains over state-of-the-art approaches in multimodal sentiment analysis and humor detection tasks. Explorative analysis reveals desirable traits, such as reductions in modality gap, being learned by the representation learning functions, which obviates the need for complex fusion mechanism. Overall, we stress the importance of representation learning as a pre-cursory step of fusion and demonstrate its efficacy through rigorous experimentation. The codes for our experiments are available at: https://github.com/declare-lab/MISA.

In the future, we plan to analyze MISA in other dimensions of affect, such as emotions. Additionally, we also aim to combine the MISA framework with other fusion schemes to try for further improvements. Finally, the similarity and difference loss modeling allow various metrics and regularization choices. We thus intend to analyze other options in this regard.

ACKNOWLEDGMENTS

This research is supported by (*i*) Singapore Ministry of Education Academic Research Fund Tier 2 under MOE's official grant number MOE2018-T2-1-103, (*ii*) A*STAR under its RIE 2020 Advanced Manufacturing and Engineering (AME) programmatic grant, Award No. - A19E2b0098, and (*iii*) DSO, National Laboratories, Singapore under grant number RTDST190702 (Complex Question Answering).

REFERENCES

[1] Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. *CoRR* abs/2004.02105 (2020). arXiv:2004.02105 https://arxiv.org/abs/2004.02105

[2] Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, MN, USA, 370–379. https://doi.org/10.18653/v1/n19-1034

[3] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013 (JMLR Workshop and Conference Proceedings, Vol. 28)*. JMLR.org, Atlanta, GA, USA, 1247–1255. http://proceedings.mlr.press/v28/andrew13.html

[4] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. IEEE Computer Society, Lake Placid, NY, USA, 1–10. https://doi.org/10.1109/WACV.2016.7477553

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain Separation Networks. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., Barcelona, Spain, 343–351. http://papers.nips.cc/paper/6254-domain-separation-networks

[6] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5647–5657. https://doi.org/10.18653/v1/D19-1566

[7] Feiyang Chen, Ziqian Luo, Yanyan Xu, and Dengfeng Ke. 2019. *Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis*. Technical Report. EasyChair.

[8] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK) *(ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 163–171. https://doi.org/10.1145/3136755.3136801

[9] Randall Davis. 2012. Multi-View Latent Variable Discriminative Models for Action Recognition. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (CVPR '12)*. IEEE Computer Society, USA, 2120–2127.

[10] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*. IEEE, Florence, Italy, 960–964. https://doi.org/10.1109/ICASSP.2014.6853739

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Thomas Drugman and Abeer Alwan. 2011. Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association*. ISCA, Florence, Italy, 1973–1976. http://www.isca-speech.org/archive/interspeech_2011/i11_1973.html

[13] Thomas Drugman, Mark R. P. Thomas, Jón Guðnason, Patrick A. Naylor, and Thierry Dutoit. 2012. Detection of Glottal Closure Instants From Speech Signals: A Quantitative Review. *IEEE Trans. Audio, Speech & Language Processing* 20, 3 (2012), 994–1006. https://doi.org/10.1109/TASL.2011.2170835

[14] Rosenberg Ekman. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

[15] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 457–468. https://doi.org/10.18653/v1/D16-1044

[16] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3454–3466. https://doi.org/10.18653/v1/D18-1382

[17] Yue Gu, Xinyu Li, Kaixiang Huang, Shiyu Fu, Kangning Yang, Shuhong Chen, Moliang Zhou, and Ivan Marsic. 2018. Human Conversation Analysis Using Attentive Multimodal Networks with Hierarchical Encoder-Decoder. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 537–545. https://doi.org/10.1145/3240508.3240714

[18] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access* 7 (2019), 63373–63394. https://doi.org/10.1109/ACCESS.2019.2916887

[19] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2046–2056. https://doi.org/10.18653/v1/D19-1211

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[21] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S. Mukherjee, Timothy M. Hospedales, Neil Martin Robertson, and Yongxin Yang. 2017. Attribute-Enhanced Face Recognition with Neural Tensor Fusion Networks. In *IEEE International Conference on Computer Vision, ICCV 2017*. IEEE Computer Society, Venice, Italy, 3764–3773. https://doi.org/10.1109/ICCV.2017.404

[22] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop*. Curran Associates, Inc., Vancouver, Canada. https://vigilworkshop.github.io/static/papers/40.pdf

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR* abs/1908.03557 (2019). arXiv:1908.03557 http://arxiv.org/abs/1908.03557

[24] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 150–161. https://doi.org/10.18653/v1/d18-1014

[25] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*. Association for Computational Linguistics, Vancouver, Canada, 1–10. https://doi.org/10.18653/v1/P17-1001

[26] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*. Association for Computational Linguistics, Melbourne, Australia, 2247–2256. https://doi.org/10.18653/v1/P18-1209

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., Vancouver, BC, Canada, 13–23. http://papers.nips.cc/paper/8297-vilbert-pretraining-task-agnostic-visiolinguistic-representations-for-vision-and-language-tasks

[28] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[29] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 481–492. https://doi.org/10.18653/v1/P19-1046

[30] Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, New York, NY, USA, 164–172. https://aaai.org/ojs/index.php/AAAI/article/view/5347

[31] Sijie Mai, Songlong Xing, and Haifeng Hu. 2020. Locally Confined Modality Fusion Network With a Global Perspective for Multimodal Human Affective Computing. *IEEE Trans. Multimedia* 22, 1 (2020), 122–137. https://doi.org/10.1109/TMM.2019.2925966

[32] Navonil Majumder, Devamanyu Hazarika, Alexander F. Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. Based Syst.* 161 (2018), 124–133. https://doi.org/10.1016/j.knosys.2018.07.041

[33] Rada Mihalcea. 2012. Multimodal Sentiment Analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis* (Jeju, Republic of Korea) *(WASSA '12)*. Association for Computational Linguistics, USA, 1.

[34] David Olson. 1977. From utterance to text: The bias of language in speech and writing. *Harvard educational review* 47, 3 (1977), 257–281.

[35] Gwangbeen Park and Woobin Im. 2016. Image-Text Multi-Modal Representation Learning by Adversarial Backpropagation. *CoRR* abs/1612.08354 (2016). arXiv:1612.08354 http://arxiv.org/abs/1612.08354

[36] Minlong Peng, Qi Zhang, Yu-Gang Jiang, and Xuanjing Huang. 2018. Cross-Domain Sentiment Classification with Target Domain Specific Information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*. Association for Computational Linguistics, Melbourne, Australia, 2505–2513. https://doi.org/10.18653/v1/P18-1233

[37] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-Modal Generative Adversarial Networks for Common Representation Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1, Article 22 (Feb. 2019), 24 pages. https://doi.org/10.1145/3284750

[38] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[39] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *The Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii, 6892–6899. https://doi.org/10.1609/aaai.v33i01.33016892

[40] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabás Poczós. 2018. Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. Association for Computational Linguistics, Melbourne, Australia, 53–63. https://doi.org/10.18653/v1/W18-3308

[41] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98 – 125. https://doi.org/10.1016/j.inffus.2017.02.003

[42] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2539–2544. https://doi.org/10.18653/v1/D15-1303

[43] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. In *2017 IEEE International Conference on Data Mining, ICDM 2017*. IEEE Computer Society, New Orleans, LA, USA, 1033–1038. https://doi.org/10.1109/ICDM.2017.134

[44] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 873–883. https://doi.org/10.18653/v1/P17-1081

[45] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *CoRR* abs/2005.00357 (2020). arXiv:2005.00357 https://arxiv.org/abs/2005.00357

[46] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrusaitis, and Roland Goecke. 2016. Extending Long Short-Term Memory for Multi-View Structured Learning. In *Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 9911)*. Springer, Amsterdam, The Netherlands, 338–353. https://doi.org/10.1007/978-3-319-46478-7_21

[47] Sebastian Ruder and Barbara Plank. 2018. Strong Baselines for Neural Semi-Supervised Learning under Domain Shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*. Association for Computational Linguistics, Melbourne, Australia, 1044–1054. https://doi.org/10.18653/v1/P18-1096

[48] Mathieu Salzmann, Carl Henrik Ek, Raquel Urtasun, and Trevor Darrell. 2010. Factorized Orthogonal Latent Spaces. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010 (JMLR Proceedings, Vol. 9)*. JMLR.org, Sardinia, Italy, 701–708. http://proceedings.mlr.press/v9/salzmann10a.html

[49] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *CoRR* abs/1908.08530 (2019). arXiv:1908.08530 http://arxiv.org/abs/1908.08530

[50] Zhongkai Sun, Prathusha K. Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning Relationships between Text, Audio, and Video via Deep Canonical Correlation for Multimodal Language Analysis. *CoRR* abs/1911.05544 (2019). arXiv:1911.05544 http://arxiv.org/abs/1911.05544

[51] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, New Orleans, LA, USA. https://openreview.net/forum?id=rygqqsA9KX

[52] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[53] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous Deep Transfer Across Domains and Tasks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. IEEE Computer Society, USA, 4068–4076. https://doi.org/10.1109/ICCV.2015.463

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., Long Beach, CA, USA, 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[55] Weiran Wang, Honglak Lee, and Karen Livescu. 2016. Deep Variational Canonical Correlation Analysis. *CoRR* abs/1610.03454 (2016). arXiv:1610.03454 http://arxiv.org/abs/1610.03454

[56] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI Press, Honolulu, Hawaii, 7216–7223. https://doi.org/10.1609/aaai.v33i01.33017216

[57] Chen Xi, Guanming Lu, and Jingjie Yan. 2020. Multimodal Sentiment Analysis Based on Multi-Head Attention Mechanism. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing* (Haiphong City, Viet Nam) *(ICMLSC 2020)*. Association for Computing Machinery, New York, NY, USA, 34–39. https://doi.org/10.1145/3380688.3380693

[58] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/D17-1115

[59] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*. AAAI Press, New Orleans, Louisiana, USA, 5634–5641. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17341

[60] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*. Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[61] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*. AAAI Press, New Orleans, Louisiana, USA, 5642–5649. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17390

[62] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88. https://doi.org/10.1109/MIS.2016.94

[63] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, Toulon, France. https://openreview.net/forum?id=SkB-_mcel

A BASELINE MODELS

This section provides the details of the baseline models mentioned in the paper.

- **MFN** [59] [5] is a multi-view gated memory network that stores intra- and cross-view interactions in its memories.
- **RAVEN** [56] utilizes an attention-based model on non-verbal signals to re-adjst word embeddings based on the multimodal context.
- **MARN** [61] learns intra-modal and cross-modal interactions by designing a hybrid LSTM memory component.
- **MV-LSTM** [46] proposes a multi-view LSTM variant with designated representations for each view inside the LSTM function.

[5] https://github.com/pliang279/MFN

- **RMFN** [24] decomposes the fusion process into multi-stage computations, where each stage focuses on a subset of multimodal signals.
- **Graph-MFN** [60] is a development over the MFN fusion model which adds a dynamic graph module on top of its recurrent structure. The nodes of the graph are the various interactions (bi-modal, tri-modal) with a hierarchical topology.
- **TFN** [58] [6] calculates a multi-dimensional tensor (based on outer-product) to capture uni-, bi-, and tri-modal interactions.
- **LMF** [26] [7] is an improvement over TFN, where low-rank modeling of the TFN tensor is proposed.
- **MFM** [51] [8] learns discriminative and generative representations for each modality where the former is used for classification while the latter is used to learn the modality-specific generative features.
- **LMFN** [31] segments the utterance vectors from each modality into smaller sections and performs fusion in these local regions, followed by global fusion across these fused segments.
- **HFFN** [29] follows a similar strategy where the local fusion is termed as *divide*, and *combine*, while the global fusion is termed as *conquer* phase.
- **MulT** [52] proposes a multimodal transformer architecture which translates one modality to another using directional pairwise cross-attention.
- **MCTN** [39] implements a translation-based model with an encoder-decoder setup to convert one modality to another. Coupled with cyclic consistency losses, the encoding representation learns informative common representations from all modalities.
- **ARGF** [30] is a recent model that learns a common embedding space by translating a source modality to a target modality through adversarial training, and maked predictions using graph-based fusion mechanism.

From the contextual-utterance family of works, we consider the following baseline models:

- **BC-LSTM** [44] is a contextual-utterance model which learns an bi-directional LSTM model overall the whole video, thus framing the problem as a structured prediction task.
- **CH-FUSION** [32] is a strong baseline which performs hierarchical fusion by composing bi-modal interactions followed by tr-modal fusion.
- **CIM-MTL** [2] is a multi-task version of MMMU-BA, analyzing the sentiment as well as emotion to exploit finds the inter-dependence between them.
- **CIA** [6] is another contextual model whose core functionality is to learn cross-modal auto-encoding to learn modality translations and utilize this feature in a contexual attention framework.
- **DFF-ATMF** [7] is a bi-modal model which first learns individual modality features followed by attention-based modality fusion.
- **C-MFN** [19] is the present state-of-the-art for UR_FUNNY. Essentially it is based on the MFN fusion mechanism and extends it to the contextual regime which takes the previous sequence of utterances into account along with MFN-style multimodal fusion.

[6] https://github.com/Justin1904/TensorFusionNetworks
[7] https://github.com/Justin1904/Low-rank-Multimodal-Fusion
[8] https://github.com/pliang279/factorized/

B DATASET SIZES

Table 5 provides the sizes (number of utterances) in each dataset.

Datasets mode	MOSI #utterances	MOSEI #utterances	UR_FUNNY #utterances
train	1283	16315	10598
dev	229	1871	2626
test	686	4654	3290

Table 5: Sizes of the datasets.

C HYPER-PARAMETER SELECTION

To select appropriate hyper-parameters, we utilize the validation sets provided in the datasets. We perform grid-search over the hyper-parameters to select the model with best validation classification/regression loss. We look over finite sets of options for hyper-parameters. These include non-linear activations: *leakyrelu*[9], *prelu*[10], *elu*[11], *relu*[12], and *tanh*[13], $\alpha \in \{0.7, 1.0\}$, $\beta \in \{0.3, 0.7, 1.0\}$, and $\gamma \in \{0.7, 1.0\}$. Finally, we look at dropout values from $\{0.1, 0.5, 0.7\}$. For optimization, we utilize the Adam optimizer with an exponential decay learning rate scheduler. The training duration of each model is governed by early-stopping strategy with a patience of 6 epochs. The final hyper-parameters for each model per dataset is summarized in Table 6.

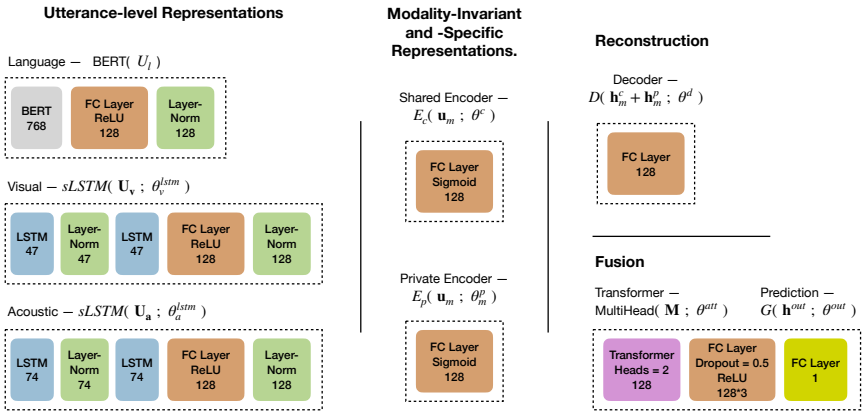
Hyper-param	MOSI	MOSEI	UR_FUNNY
cmd K	5	5	5
activation	ReLU	LeakyReLU	Tanh
batch size	64	16	32
gradient clip	1.0	1.0	1.0
α	1.0	0.7	0.7
β	0.3	0.3	1.0
γ	1.0	0.7	1.0
dropout	0.5	0.1	0.1
d_h	128	128	128
learning rate	1e-4	1e-4	1e-4

Table 6: Final hyper-parameter values in each dataset.

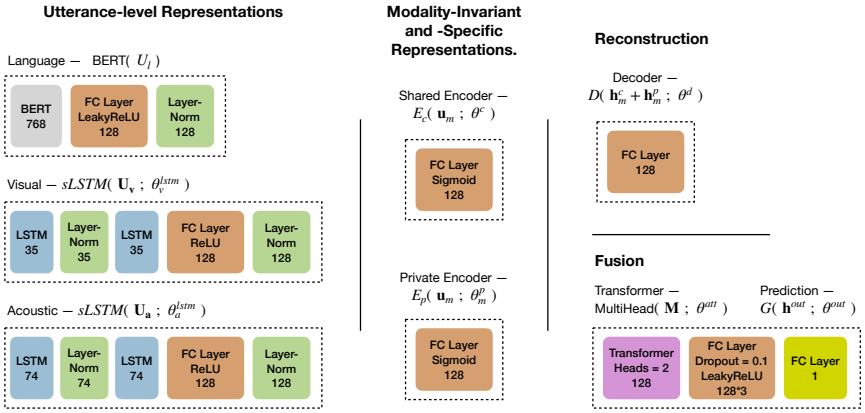
D NETWORK TOPOLOGY

Fig. 6 describes the network topologies of the final models used in each dataset.

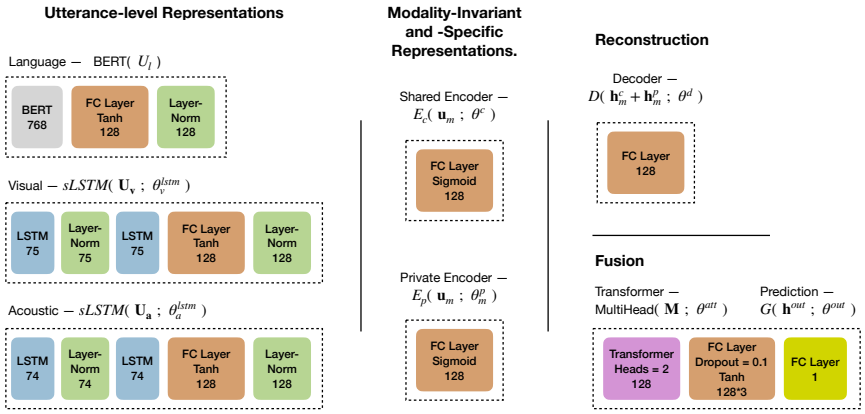
[9] https://pytorch.org/docs/stable/nn.html#leakyrelu
[10] https://pytorch.org/docs/stable/nn.html#prelu
[11] https://pytorch.org/docs/stable/nn.html#elu
[12] https://pytorch.org/docs/stable/nn.html#relu
[13] https://pytorch.org/docs/stable/nn.html#tanh



(a) MOSI



(b) MOSEI



(c) UR_FUNNY

Figure 6: Description of the topologies used for the different datasets.