Self-Supervised Unimodal Label Generation Strategy Using Recalibrated Modality Representations for Multimodal Sentiment Analysis

Yewon Hwang and **Jong-Hwan Kim**
School of Electrical Engineering, KAIST
Daejeon, Republic of Korea
{ywhwang, johkim}@rit.kaist.ac.kr

Abstract

While multimodal sentiment analysis (MSA) has gained much attention over the last few years, the main focus of most work on MSA has been limited to constructing multimodal representations that capture interactions between different modalities in a single task. This was largely due to a lack of unimodal annotations in MSA benchmark datasets. However, training a model using only multimodal representations can lead to suboptimal performance due to insufficient learning of each uni-modal representation. In this work, to fully optimize learning representations from multimodal data, we propose SUGRM which jointly trains multimodal and unimodal tasks using recalibrated features. The features are recalibrated such that the model learns to weight the features differently based on the features of other modalities. Further, to leverage unimodal tasks, we auto-generate unimodal annotations via a unimodal label generation module (ULGM). The experiment results on two benchmark datasets demonstrate the efficacy of our framework. [1]

1 Introduction

These days, we can easily spot AI systems in our society that serve or assist humans. Understanding human emotions has become a critical factor for these AI systems to seamlessly integrate into human's life (Castillo et al., 2018; De Graaf and Allouch, 2013). However, understanding humans' emotions is not a trivial task. This is because humans tend to express their feelings through multiple cues in a complex form. Emotions can be expressed simply through language, but they can also be manifested through facial expression, behaviors or even tone of voice (Morency et al., 2011). Moreover, sometimes these cues signal a compatible emotion, while other times they signal conflicting emotions, e.g., positive language with a condescending tone of voice indicates sarcasm (Robins et al., 2009).

Taking this nature into account, multimodal sentiment analysis (MSA) has become an active field of research which aims to understand the affective state of humans through visual, acoustic, and textual features. In general, when working with multimodal data like in MSA, each modality contains both supplementary and complementary information to each other, providing richer information about the data. This leads to improved performance over using only one modality (Vaezi Joze et al., 2020). However, capturing information in each modality as well as modeling the interactions between different modalities still remain challenging tasks to unravel (Hazarika et al., 2020).

Most of the existing works on MSA revolve around learning a joint representation which encompasses information from all modalities through sophisticated fusion methods varying from tensor-based (Zadeh et al., 2017) to attention-based methods (Tsai et al., 2019; Rahman et al., 2020), where the learning process happens in a single task. Single task learning was a dominant learning framework in MSA particularly due to the nature of the benchmark datasets: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Bagher Zadeh et al., 2018). Considering all modalities, only one comprehensive sentiment intensity value (i.e., multimodal label, y_m) is annotated in both datasets due to the laborious labeling process. Meaning, unimodal labels (y_t, y_a, y_v) are omitted in the datasets. However, a recent study (Yu et al., 2021) argued the absence of unimodal annotations hinders capturing modality-specific information and proposed a module that auto-generates unimodal annotations from the multimodal labels.

In this work, we propose a novel framework, SUGRM, which leverages a self-supervised unimodal label generation strategy using recalibrated modality representations for MSA. First, we recali-

[1] Our code is available at: https://github.com/skystarhyw/SUGRM

Findings of the Association for Computational Linguistics: EACL 2023, pages 35–46

brate modality representations using Modality Recalibration Module (MRM). This allows the model to dynamically adjust features based on the features of other modalities. Further, motivated by (Yu et al., 2021), we propose a new unimodal label generation module (ULGM), which generates unimodal annotations (y_t, y_a, y_v) based on the multimodal annotation (y_m) in a self-supervised manner.

Different from (Yu et al., 2021), which preserves feature space of each modality, we project features of each modality into a common semantic feature space. Thus, our ULGM hypothesizes the distance between two features in a common semantic feature space is proportional to the distance between the corresponding labels in a label space. This not only allows simpler calculation of the offset (see section 3.3), but also avoids the problem in (Yu et al., 2021); that is, when two distances from a multimodal feature 1) to the center of negative multimodal features and 2) to the center of positive multimodal features are approximately equal, the generated unimodal label diverges. This could lead to unstable learning, potentially causing the model to fall into a local minima.

Our experiment results not only empirically validate our hypothesis, but also prove that using recalibrated modality representation as well as our ULGM lead to enhanced performance. The main contributions of our work can be summarized as follows:

- We introduce Modality Recalibration Module (MRM) for MSA which recalibrates modality features based on features of other modalities.
- We design a novel unimodal label generation module (ULGM) to expand MSA to multi-task learning and jointly train unimodal and multimodal tasks.
- Not only does our method outperform the previous SOTA results, but the experiment results validate the effectiveness of our framework.

2 Related Work

Prior works of MSA mainly focused on improving fusion between multi-modalities as well as learning joint representations. In earlier works, early fusion (Pérez-Rosas et al., 2013; Poria et al., 2016) and late fusion (Zadeh et al., 2016) were popular fusion methods to combine the multiple modalities. Later, more sophisticated methods of fusion were proposed using a multi-dimensional tensor (Zadeh et al., 2017), attention mechanism (Zadeh et al., 2018a,b), multi-stage fusion (Liang et al., 2018) and low rank tensors to improve efficiency of fusion (Liu et al., 2018). In (Wang et al., 2019), the authors dynamically adjusted a word representation by calculating a shift caused by accompanying nonverbal information. More recent works have focused on applying Transformer architecture to better capture interactions between modalities and learn feature representations. For instance, (Rahman et al., 2020) was directly built upon (Wang et al., 2019), but used pretrained Transformer based language models to improve the performance. (Tsai et al., 2019) proposed cross-modal attention to latently adapt a target modality from source modalities. (Cheng et al., 2021) reduced the computational burden in (Tsai et al., 2019), by generating sparse attention matrices and compressing a long sequence to a short sequence. Further, a multi-task learning approach has been applied in recent MSA (Akhtar et al., 2019; Yu et al., 2021) to increase data efficiency.

Taking inspiration from the previous work (Yu et al., 2021), we expand a learning framework of MSA to multi-task learning. The benefits of multi-task learning is that each task helps a learning process of other tasks. This allows the model to learn better generalized representations that are shared across the tasks. Further, we recalibrate features of each modality and efficiently model inter-, intra-modality relationships by adopting the work of (Hu et al., 2018; Vaezi Joze et al., 2020; Cheng et al., 2021).

3 Methodology

3.1 Problem Definition

We define the input to the model as $I_{s\in\{t,a,v\}}$ which is composed of three types of modalities-text, audio, and video. The goal of our model is to take I_s as input and predict a sentiment intensity $\hat{y} \in \mathbb{R}$. To aid the learning process, our model generates labels for each modality $y_s \in \mathbb{R}$ during training.

3.2 Overall Architecture

Our framework consists of multimodal and unimodal tasks where they share modality representations as shown in Figure 1.

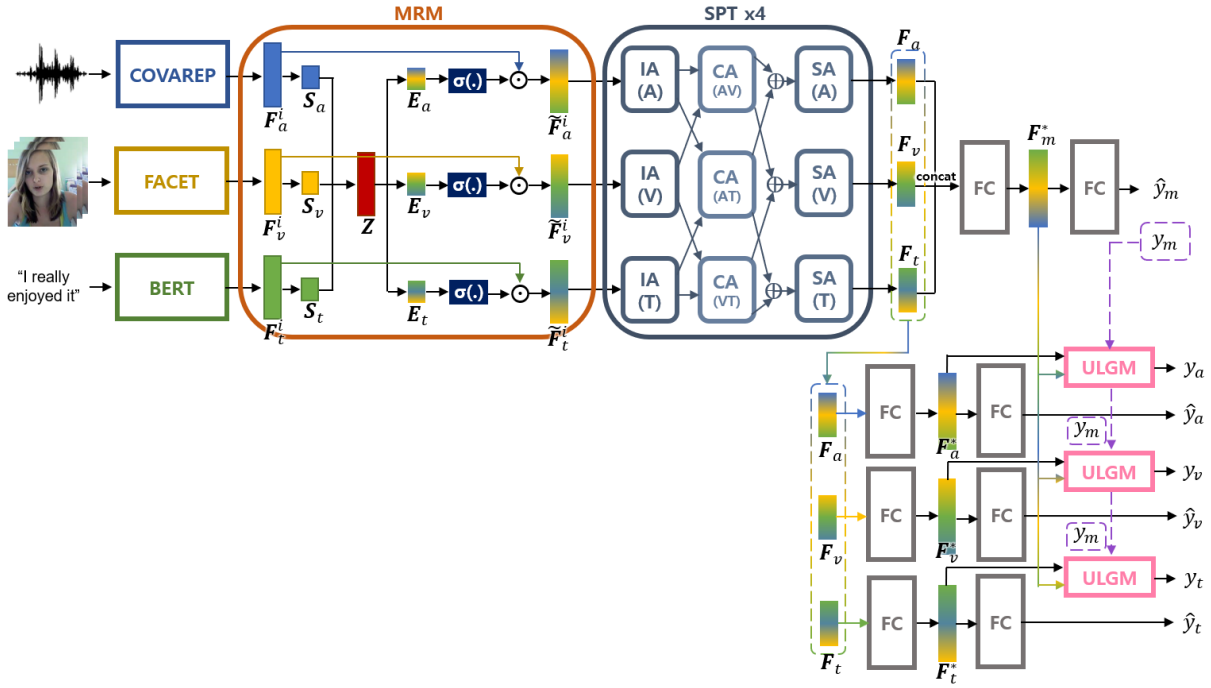

Figure 1: The overall architecture of SUGRM. The y_a, y_v, and y_t are the unimodal annotations generated from our ULGM based on the human-annotated multimodal label y_m to enable supervised learning of the unimodal tasks. The $\hat{y}_a$, $\hat{y}_v$, $\hat{y}_t$, and $\hat{y}_m$ are the predicted sentiment values from the unimodal and multimodal tasks.

3.2.1 Multimodal Task

In the multimodal task, modality features ($\boldsymbol{F}^i_{s\in\{t,a,v\}}$) are initially extracted from pretrained BERT (Devlin et al., 2019), COVAREP (Degottex et al., 2014), and FACET (iMotions, 2013) for textual, acoustic, and visual information, respectively. Then these features are passed through Modality Recalibration Module (MRM) for feature recalibration. After the features are recalibrated, the final feature representation of each modality is captured using Sparse Phased Transformer (SPT).

Modality Recalibration Module. MRM recalibrates modality features using squeeze and excitation (SE) technique (Hu et al., 2018). This particular idea was studied in the case of CNN in (Vaezi Joze et al., 2020). Here, we show how SE can be expanded to the MSA application. MRM receives $\boldsymbol{F}^i_s \in \mathbb{R}^{l_s \times d_s}$ as input, where l_s is the sequence length and d_s is the feature dimension of s-modality, and squeeze the input along the sequence length using global average pooling:

$$\boldsymbol{S}_s(d) = \frac{1}{l_s}\sum_{l=1}^{l_s} \boldsymbol{F}^i_s(l,d)\,,$$

where $s \in \{t,a,v\}$ and $d = 1,...,d_s$. Then the excitation process is performed to apply different weight calibrations for each modality. First, squeezed features are concatenated and fed into a series of a fully connected network and ReLU to learn a global multimodal embedding $\boldsymbol{Z}$:

$$\boldsymbol{Z} = ReLU(\boldsymbol{W}_z[\boldsymbol{S}_t;\boldsymbol{S}_a;\boldsymbol{S}_v] + \boldsymbol{b}_z)\,.$$

Here, the fully connected network reduces feature dimension. Then we compute excitation signals using another fully connected network as follows:

$$\boldsymbol{E}_s = \boldsymbol{W}_s\boldsymbol{Z} + \boldsymbol{b}_s\,.$$

The second fully connected network restores the original feature dimension, adopting bottleneck architecture. The reason for this is to reduce the number of computations and improve generalization (Hu et al., 2018). Finally, the input features are recalibrated through a following gating mechanism:

$$\tilde{\boldsymbol{F}}^i_s = 2\times\sigma(\boldsymbol{E}_s)\odot\boldsymbol{F}^i_s\,,$$

where $\sigma(\cdot)$ is the sigmoid function and $\odot$ is the element-wise product along the feature dimension. Since the numbers returned by sigmoid function (between 0 and 1) are multiplied by the original features, each feature is rescaled based on its importance. Finally, the textual, acoustic, and visual features after MRM can be described as follows:

$$\tilde{\boldsymbol{F}}^i_s = MRM(\boldsymbol{F}^i_s;\theta^{mrm}) \in \mathbb{R}^{l_s\times d_s}\,,$$

where θ^{mrm} are the parameters of MRM.

Sparse Phased Transformer. SPT (Cheng et al., 2021) extracts the final feature representation of each modality using the recalibrated features. The motivation behind SPT is twofold: to extract more informative features by modeling intra- and inter-modalities (preferred over LSTM[2]) and to build a more efficient and lighter model (preferred over (Cheng et al., 2021)[2]). SPT alleviates the computational burden of the self-attention mechanism in the vanilla Transformer. Instead of generating a full attention matrix, SPT generates a sparse attention matrix to reduce computational complexity.[3] Multimodal SPT is composed of input attention, cross attention, and self attention. Input attention (IA) compresses input sequence into hidden states. Then the hidden states of two different modalities are interacted through cross attention (CA). Finally, self attention (SA) refines the feature representations of each modality. For the technical details of SPT, refer to (Cheng et al., 2021) on which our implementation of SPT is based.

We denote the final feature representation for each modality as follows:

$$\boldsymbol{F}_s = SPT(\tilde{\boldsymbol{F}}_s^i; \theta^{spt}) \in \mathbb{R}^{d_s} ,$$

where SPT is the process of [IA→CA→SA] repeated 4 times and θ^{spt} are the parameters of SPT. Finally, the last element of the sequence is selected as a sequence representation.

To obtain a fusion representation, we concatenate each modality representation and project into a lower-dimensional feature space $\mathbb{R}^{d_c}$ as follows:

$$\boldsymbol{F}_m^* = ReLU(\boldsymbol{W}_1^m[\boldsymbol{F}_t; \boldsymbol{F}_a; \boldsymbol{F}_v] + \boldsymbol{b}_1^m) .$$

Lastly, the multimodal sentiment is predicted as follows:

$$\hat{y}_m = \boldsymbol{W}_2^m \boldsymbol{F}_m^* + b_2^m .$$

3.2.2 Unimodal Task

For the unimodal task, we use the feature representation of each modality obtained from the multimodal task ($\boldsymbol{F}_{s\in\{t,a,v\}}$). Then we map each feature representation into the same feature space as $\mathbb{R}^{d_c}$ (i.e., a common semantic feature space) as follows:

$$\boldsymbol{F}_s^* = ReLU(\boldsymbol{W}_1^s \boldsymbol{F}_s + \boldsymbol{b}_1^s) .$$

[2]Three options were considered as a final feature extractor: LSTM, multimodal Transformer (Cheng et al., 2021), and SPT (See Table 4).

[3]The authors of SPT (Cheng et al., 2021) claim that the number of parameters is reduced to 10% of (Tsai et al., 2019) which utilizes the vanilla Transformer encoder.

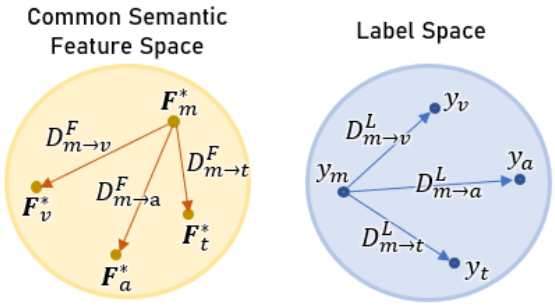

Figure 2: The distance from multimodal feature ($\boldsymbol{F}_m^*$) to s-modal feature ($\boldsymbol{F}_s^*$) in a common semantic feature space: $D_{m\to s}^F$, and the distance from multimodal label (y_m) to s-modal label (y_s) in a label space: $D_{m\to s}^L$.

Then the final sentiment prediction from each modality is obtained through an independent fully-connected layer:

$$\hat{y}_s = \boldsymbol{W}_2^s \boldsymbol{F}_s^* + b_2^s .$$

The unimodal tasks are trained using supervised learning, where labels for each modality are obtained via non-parametric Unimodal Label Generation Module (ULGM):

$$y_s = ULGM(y_m, \boldsymbol{F}_m^*, \boldsymbol{F}_s^*) .$$

Finally, the multimodal task and three unimodal tasks are jointly trained.

3.3 ULGM

The goal of ULGM is to generate labels for each unimodality based on multimodal labels and modality representations. As shown in Figure 2, our ULGM is designed based on the notion that the distance between two features in a common semantic feature space is proportional to the distance between the corresponding labels in a label space:

$$D_{m\to s}^F \propto D_{m\to s}^L ,$$

where $s \in \{t, a, v\}$. Our ULGM computes the offset of unimodal label y_s with respect to the multimodal label y_m based on the distance from the multimodal feature to each unimodal feature. We consider two factors when computing the offset: the magnitude and the direction.

Magnitude of offset. To calculate the offset, we argue that the maximum distance within the common semantic feature space is proportional to the maximum distance within the label space. In CMU-MOSI and -MOSEI datasets, the multimodal

labels range from -3 to +3, meaning the distance between multimodal features with labels -3 ($\boldsymbol{F}_m^{*-3}$) and +3 ($\boldsymbol{F}_m^{*+3}$) must correspond to the maximum distance within the common semantic feature space. Therefore, any $D^F_{m\to s}$ greater than the maximum distance is clipped to $D^F_{max} = ||\overline{\boldsymbol{F}_m^{*+3}} - \overline{\boldsymbol{F}_m^{*-3}}||$:

$$D^F_{m\to s} = \begin{cases} ||\boldsymbol{F}^*_m - \boldsymbol{F}^*_s||, & \text{if } \mathrm{D}^F_{m\to s} \le D^F_{max}\,, \\ D^F_{max}, & \text{otherwise}\,, \end{cases}$$

where $\overline{\boldsymbol{F}_m^{*+3}}$ and $\overline{\boldsymbol{F}_m^{*-3}}$ are the mean of $\boldsymbol{F}_m^{*+3}$ and $\boldsymbol{F}_m^{*-3}$, respectively, and $||\cdot||$ is $L2$ normalization.

Based on our notion and the above argument, we can consider the following relationship from which we can obtain the magnitude of the offset from a multimodal label to an unimodal label:

$$D^F_{m\to s}/D^F_{max} = D^L_{m\to s}/D^L_{-3\to+3}\,,$$

$$D^L_{m\to s} = \frac{D^F_{m\to s}}{D^F_{max}} D^L_{-3\to+3}\,.$$

Direction of offset. In order to determine the direction of the offset, we identify the position of the s-modal feature with respect to the multimodal feature. To do that, we first take the average of the multimodal features with positive annotations ($\overline{\boldsymbol{F}_m^{*+}}$) and negative annotations ($\overline{\boldsymbol{F}_m^{*-}}$). Then we locate the multimodal and the s-modal features within this realm of feature space as shown in Figure 3. Using the distance from modality representations ($\boldsymbol{F}^*_{x\in\{m,t,a,v\}}$) to $\overline{\boldsymbol{F}_m^{*+}}$ and $\overline{\boldsymbol{F}_m^{*-}}$, we can determine the direction of the offset as follows:

$$Direction = \begin{cases} +, & \text{if } \frac{D^p_s}{D^n_s} < \frac{D^p_m}{D^n_m}\,, \\ -, & \text{if } \frac{D^p_s}{D^n_s} > \frac{D^p_m}{D^n_m}\,, \\ 0, & \text{if } \frac{D^p_s}{D^n_s} = \frac{D^p_m}{D^n_m}\,, \end{cases}$$

where $D^p_s = ||\boldsymbol{F}^*_s - \overline{\boldsymbol{F}_m^{*+}}||$, $D^n_s = ||\boldsymbol{F}^*_s - \overline{\boldsymbol{F}_m^{*-}}||$, $D^p_m = ||\boldsymbol{F}^*_m - \overline{\boldsymbol{F}_m^{*+}}||$, and $D^n_m = ||\boldsymbol{F}^*_m - \overline{\boldsymbol{F}_m^{*-}}||$. Finally, we obtain the unimodal label y_s as follows:

$$y_s = \begin{cases} y_m + D^L_{m\to s}, & \text{if direction is } +\,, \\ y_m - D^L_{m\to s}, & \text{if direction is } -\,, \\ y_m, & \text{if direction is } 0\,. \end{cases}$$

Unimodal Label Update Scheme. We update the generated unimodal labels using a momentum-based update policy (Yu et al., 2021) as follows:

$$y^e_s = \begin{cases} y_m & \text{for } e = 1\,, \\ \frac{e-1}{e+1} y^{(e-1)}_s + \frac{2}{e+1} y^e_s & \text{for } e > 1\,, \end{cases}$$

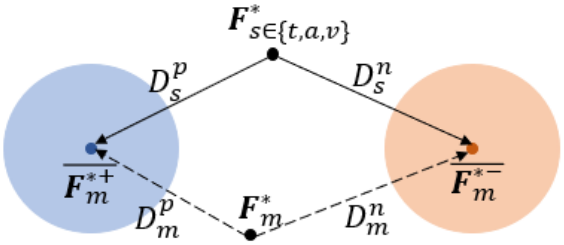

Figure 3: An illustration of positions of modality representations with respect to the mean of multimodal representations with positive labels ($\overline{\boldsymbol{F}_m^{*+}}$) and negative labels ($\overline{\boldsymbol{F}_m^{*-}}$) in the common semantic feature space.

where $s \in \{t, a, v\}$ and e is epoch. This scheme is used to mitigate the instability of labels that are generated at the beginning of epochs in which the learning of the modality features is trivial. This update scheme allows the labels generated in later epochs to have greater impact than the ones generated in earlier epochs. After a sufficient number of iterations, unimodal labels become stabilized, resulting in a stable training process of unimodal tasks. As can be seen in Figure 4, the labels stabilize within 15 epochs.

3.4 Objective Function for Training

For the objective function, we investigated three loss functions that are widely used in regression tasks: $L1$ loss, $L2$ loss, and Huber loss. Based on our loss ablation study (see Table 8 in Appendix), we use $L1$ loss as the objective function for both multimodal and unimodal tasks. We minimize the sum of the two loss functions over N training samples to optimize the entire model as follows:

$$L = \frac{1}{N}\sum_i^N (|\hat{y}^i_m - y^i_m| + \sum_s^{\{t,a,v\}} w^i_s * |\hat{y}^i_s - y^i_s|)\,,$$

where the first term corresponds to the multimodal task, and the second term corresponds to the unimodal tasks optimization. Note the loss functions for the unimodal tasks are weighted by w^i_s, where $w^i_s = tanh(|y^i_s - y^i_m|)$ (Yu et al., 2021) such that the model can target the samples with larger difference between the multimodal label and the generated unimodal label more rigorously during training.

4 Experimental Settings

4.1 Datasets

We use the two most popular English benchmark datasets for MSA: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Bagher Zadeh et al., 2018). CMU-MOSI dataset consists of 2,199 labeled video clips taken from 93 videos by 89 speakers. The videos were crawled from YouTube and encompass opinions on movies, books, and products. Each video is annotated with sentiment on a [-3,3] range. CMU-MOSEI dataset is the most comprehensive dataset for sentiment analysis and emotion recognition which comprises more than 65 hours worth of 23,453 annotated video segments from 1,000 speakers addressing 250 different topics. Each video is annotated with sentiment on a [-3,3] range as well as six discrete emotions: happy, sadness, anger, disgust, surprise, and fear. We only utilize sentiment values from CMU-MOSEI in this task. See Table 6 in Appendix for the dataset split.

4.2 Baselines

We compare the performance of our model with previous state-of-the-art MSA models. The superscript A indicates the proposed method only works on the aligned settings, while UA indicates the proposed method works on both unaligned and aligned settings.[4]

EF-LSTM.^{A} Early Fusion LSTM concatenates the multimodal features at the input level.

LF-LSTM.^{UA} Late Fusion LSTM combines modality-wise decisions using a voting mechanism.

TFN.^{A} The Tensor Fusion Network (Zadeh et al., 2017) models intra- and inter-modality dynamics through multi-dimensional tensors.

RAVEN.^{A} The Recurrent Attended Variation Embedding Network (Wang et al., 2019) models nonverbal sequences and dynamically shifts word representations based on nonverbal cues.

MCTN.^{A} The Multimodal Cyclic Translation Network (Pham et al., 2019) learns robust joint representations via multimodal cyclic translations using a cycle consistency loss.

[4]Multimodal data in CMU-MOSI and MOSEI are loaded from different sources which come at different frequencies, making the multimodal data "unaligned" in terms of sequence length. (The lengths of text, audio, video segments are 50, 375, 500, respectively for the unaligned dataset.) These unaligned data have been preprocessed through CMU-Multimodal SDK (https://github.com/A2Zadeh/CMU-MultimodalSDK) to align different modalities such that they have the same sequence length of 50. Note, our method works on both aligned and unaligned settings.

MulT.^{UA} The Multimodal Transformer (Tsai et al., 2019) uses cross-modal attention to model interactions between asynchronous modalities and latently adapt one modality to another.

MAG-BERT.^{A} The Multimodal Adaptation Gate for BERT (Rahman et al., 2020) is an improvement of RAVEN which applies multimodal adaptation gate at the first layer of the BERT model.

SPT.^{UA} The multimodal Sparse Phased Transformer (Cheng et al., 2021) is an improvement of MulT in terms of efficiency by using a sampling function to generate a sparse attention matrix.

Self-MM.^{UA} The Self-Supervised Multi-task Multimodal sentiment analysis network (Yu et al., 2021) generates a unimodal label for each modality and jointly trains multimodal and unimodal tasks.

4.3 Implementation Details

We trained our framework using NVIDIA TITAN Xp and Intel i7-9700K. We use the batch size of 32 and Adam as the optimizer for both datasets. For more implementation details such as hyperparameters for each dataset, see Table 7 in Appendix.

4.4 Evaluation Metrics

We evaluate our model using four metrics: weighted binary F1 score (F1-Score), binary classification accuracy (Acc_2), Mean Absolute Error (MAE), and Pearson correlation (Corr). For F1-Score and Acc_2, we report the model performance in two ways: negative/non-negative (Zadeh et al., 2017) and negative/positive (Tsai et al., 2019).

5 Results and Analysis

5.1 Quantitative Results

Tables 1 and 2 show the experiment results on the aligned and unaligned MOSI and MOSEI datasets, respectively. Our model outperformed all of the previous SOTA baseline models on all metrics for the MOSI dataset, and achieved either SOTA or comparable-to-SOTA results on the MOSEI dataset for both the aligned and unaligned datasets. Note, CTC (Graves et al., 2006) was introduced to allow some models (Wang et al., 2019; Pham et al., 2019) that originally only work on the aligned dataset to work on the unaligned dataset in Table 2. Unlike the previous observation (Tsai et al., 2019), our model shows greater strength in the unaligned dataset than the aligned dataset. This is beneficial in that it allows omission of extra data alignment

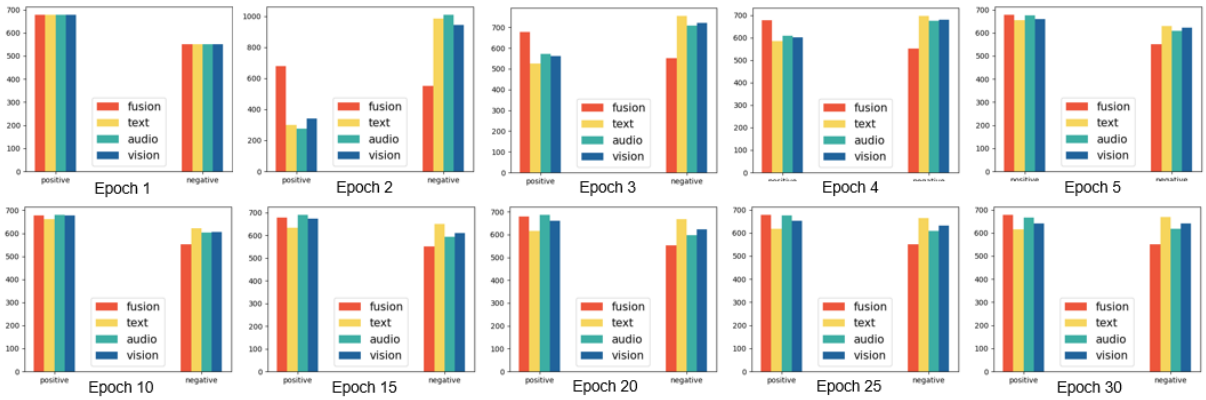

Figure 4: Visualization of the generated unimodal labels update process throughout epochs on CMU-MOSI dataset

step and data to have its inherent trait of unalignment, which could further facilitate real-time sentiment analysis.

5.2 Ablation Study

To explore the contributions of the unimodal tasks in our model, we conducted experiments using combinations of different unimodal tasks as shown in Table 3. The general trend of the results shows that incorporating the unimodal tasks leads to improvement in the model performance, which proves the effectiveness of our model. Particularly, using all three unimodal tasks along with the multimodal task resulted in substantial performance gain on all metrics compared to using the multimodal task alone on the MOSI dataset. An interesting trend on the MOSI dataset is that the performance rather decreased when only one of the unimodal tasks was added. However, we can observe that the addition of more than one unimodal task helps the model to achieve better results. On the other hand, introducing all the unimodal tasks (M,T,A,V) on the MOSEI dataset did not show as apparent performance gain as the MOSI dataset. However, we can easily observe a generally increasing trend in performance with the addition of unimodal tasks on the MOSEI dataset.

To compare our ULGM as well as the effectiveness of our architecture against that of Self-MM (Yu et al., 2021), we conducted an ablation study as shown in Table 4. Our model surpassed the performance of Self-MM via the combination of MRM, SPT, and $\text{ULGM}_{\text{ours}}$ modules. To study the effectiveness of each module, we added MRM to Self-MM, replaced LSTM in Self-MM with SPT for learning sequence representation, and replaced $\text{ULGM}_{\text{Self-MM}}$ with $\text{ULGM}_{\text{ours}}$. The addition of MRM and the replacement of SPT on the MOSI dataset certainly led to improved performance but on a limited range of metrics. However, the replacement of $\text{ULGM}_{\text{ours}}$ significantly increased the performance on all metrics. Results on the MOSEI dataset show a notable performance boost in all tasks across a wide range of metrics. Particularly, the replacement of SPT, which showed trivial results on the MOSI dataset, played an important role in improving the performance on the MOSEI dataset.

Similarly, we removed or replaced MRM, SPT, and $\text{ULGM}_{\text{ours}}$ to evaluate their contribution to our model. First, we removed MRM, replaced SPT with the vanilla Transformer encoder (TE) (Tsai et al., 2019) and LSTM, and replaced $\text{ULGM}_{\text{ours}}$ with $\text{ULGM}_{\text{Self-MM}}$. The results in Table 4 predominantly show that the inclusion of all modules results in the best performance. Replacing SPT with the vanilla Transformer encoder and $\text{ULGM}_{\text{ours}}$ with $\text{ULGM}_{\text{Self-MM}}$ led to an increase in certain metrics. However, not only the improvement is minuscule for both replacements, but the opportunity cost for exchanging computational efficiency with such minuscule improvement is rather counterproductive particularly for the SPT $\rightarrow$TE replacement.

5.3 Qualitative Results

To evaluate the quality of the generated labels of each modality, we display four samples from the CMU-MOSI dataset in Table 5. We observe that the generated unimodal annotations are generally in line with the descriptions from the text, acoustic, and visual information. This further confirms the efficacy of our ULGM.

Table 1: Results on the aligned CMU-MOSI and CMU-MOSEI datasets. In Acc_2 and F1-Score, the left side of the "/" is the "negative/non-negative" method and the right side is the "negative/positive" method.

Model	MOSI				MOSEI			
	F1-Score	Acc_2	MAE	Corr	F1-Score	Acc_2	MAE	Corr
EF-LSTM	-/75.6	-/75.8	1.053	0.613	-/78.8	-/79.1	0.665	0.621
LF-LSTM	-/75.4	-/76.4	1.037	0.620	-/80.0	-/79.4	0.625	0.655
TFN	74.1/75.2	74.8/76.0	0.955	0.649	-	-	-	-
RAVEN	-/76.6	-/78.0	0.915	0.691	-/79.5	-/79.1	0.614	0.662
MCTN	-/79.1	-/79.3	0.909	0.676	-/80.6	-/79.8	0.609	0.670
MulT	-/82.8	-/83.0	0.871	0.698	-/82.3	-/82.5	0.580	0.703
SPT	-/82.9	-/82.8	-	-	-/82.8	-/82.6	-	-
MAG-BERT	82.4/84.0	82.5/84.0	0.778	0.766	81.7/84.7	81.3/84.8	0.567	0.742
Self-MM	82.3/84.4	82.4/**84.5**	0.736	0.786	83.2/85.0	82.9/84.8	**0.533**	**0.766**
Ours	**82.8/84.5**	**82.8/84.5**	**0.723**	**0.798**	**83.9/85.1**	**83.9/85.0**	0.541	0.758

Table 2: Results on the unaligned CMU-MOSI and CMU-MOSEI datasets. Note that CTC method (Graves et al., 2006) was employed to EF-LSTM, RAVEN, and MCTN to apply these models on the unaligned setting.

Model	MOSI				MOSEI			
	F1-Score	Acc_2	MAE	Corr	F1-Score	Acc_2	MAE	Corr
EF-LSTM+CTC	-/74.5	-/73.6	1.078	0.542	-/75.9	-/76.1	0.680	0.585
LF-LSTM	-/77.8	-/77.6	0.988	0.624	-/78.2	-/77.5	0.624	0.656
RAVEN+CTC	-/73.1	-/72.7	1.076	0.544	-/75.7	-/75.4	0.664	0.599
MCTN+CTC	-/76.4	-/75.9	0.991	0.613	-/79.7	-/79.3	0.631	0.645
MulT	-/81.0	-/81.1	0.889	0.686	-/81.6	-/81.6	0.591	0.694
SPT	-/81.3	-/81.2	-	-	-/82.7	-/82.4	-	-
Self-MM	82.8/84.6	82.9/84.6	0.733	0.780	82.0/**84.6**	81.7/**84.7**	**0.530**	**0.765**
Ours	**84.3/86.3**	**84.4/86.3**	**0.703**	**0.800**	**83.6**/84.0	**83.7**/84.4	0.544	0.748

Table 3: An ablation study on the benefits of the unimodal tasks using the unaligned datasets. The bold numbers indicate the best performance, and the underlined numbers indicate enhanced performance from introducing the unimodal tasks to the multimodal task.

Model	MOSI				MOSEI			
	F1-Score	Acc_2	MAE	Corr	F1-Score	Acc_2	MAE	Corr
M	82.5/84.1	82.5/84.0	0.755	0.779	81.5/84.7	80.9/84.7	0.539	0.759
M,V	81.1/82.1	81.1/82.0	0.774	0.757	79.5/83.7	78.9/83.6	0.543	0.752
M,A	81.9/83.6	81.9/83.5	0.764	0.770	82.7/**85.2**	82.4/**85.3**	0.532	**0.763**
M,T	81.0/81.5	80.9/81.4	0.773	0.779	80.8/83.7	80.4/83.8	**0.530**	**0.763**
M,A,V	83.6/85.0	83.5/84.9	0.731	0.782	81.6/84.4	83.3/84.6	0.533	0.757
M,A,T	82.7/84.2	82.7/84.2	0.804	0.762	82.9/84.5	82.8/84.8	0.535	0.752
M,V,T	83.6/84.7	83.5/84.6	0.748	0.778	82.9/82.7	83.4/83.4	0.540	0.748
M,T,A,V	**84.3/86.3**	**84.4/86.3**	**0.703**	**0.800**	**83.6**/84.0	**83.7**/84.4	0.544	0.748

Table 4: An ablation study on the contribution of MRM, SPT, and our ULGM using the unaligned datasets. The bold numbers indicate the best performance, and the underlined numbers indicate enhanced performance compared to the baseline model. Superscript A, RP, and RM indicate added, replaced, and removed module, respectively.

Baseline	Added/Removed/ Replaced Module	MOSI				MOSEI			
		F1-Score	Acc_2	MAE	Corr	F1-Score	Acc_2	MAE	Corr
Self-MM	-	82.8/84.6	82.9/84.6	0.733	0.780	82.0/84.6	81.7/84.7	0.530	0.765
	MRM^{A}	82.4/84.1	82.5/84.2	0.718	**0.791**	**83.5**/85.0	**83.3**/85.1	0.542	0.756
	SPT^{RP}	82.8/84.3	82.8/84.3	0.735	0.785	82.7/**85.6**	82.3/**85.7**	0.534	**0.771**
	$ULGM_{ours}{}^{RP}$	**83.5/85.6**	**83.7/85.7**	**0.710**	0.790	83.0/85.3	82.7/85.3	0.538	0.757
Ours	-	**84.3/86.3**	**84.4/86.3**	**0.703**	0.800	**83.6**/84.0	83.7/**84.4**	0.544	0.748
	MRM^{RM}	81.5/82.9	81.5/82.8	0.761	0.767	79.2/83.4	78.5/83.4	**0.541**	0.746
	TE^{RP}	84.2/85.6	84.1/85.5	0.720	**0.802**	82.1/81.9	**83.8**/82.8	0.553	0.750
	$LSTM^{RP}$	79.2/81.7	79.5/81.9	0.801	0.740	77.3/82.1	76.5/82.0	0.556	0.744
	$ULGM_{Self\text{-}MM}{}^{RP}$	82.1/83.3	82.1/83.2	0.726	0.797	0.79.5/**84.1**	78.8/84.0	**0.541**	**0.756**

Table 5: Four samples from the CMU-MOSI dataset. It shows the predictions from each modality as well as the generated unimodal annotations (S_G, where S $\in \{T, A, V\}$) during training.

Text	Acoustic	Visual	Prediction	Annotation
"Everytime that was like a jump everyone jumped,"	Fast paced slightly thrilled	slightly smiling	M: 0.1, T: 0.1 A: 0.5, V: 0.7	M: 0.8, T_G: 0.6 A_G: 0.9, V_G: 0.7
"I was really hoping that this one be just as good."	Monotonic and emphasis on "really"	Slightly frowning	M: -0.1, T: -0.2 A: 0.5, V: 0.2	M: -0.8, T_G: -0.3 A_G: 0.0, V_G: -0.7
"Looks exactly the same as this character in Defiance."	Relaxed and firm	Squinting eye and raising eyebrows	M: 0.2, T: -0.1 A: 0.5, V: 0.3	M: 0.2, T_G: 0.1 A_G: 0.7, V_G: 0.1
"I don't know what they are complaining about it."	High pitched and emphasis on "what"	smiling and head roll on "what"	M: 1.1, T: 0.3 A: 1.7, V: 1.6	M: 1.8, T_G: 0.9 A_G: 1.5, V_G: 1.5

6 Conclusion and Future Work

In this paper, we proposed SUGRM, a novel framework for multimodal sentiment analysis (MSA) which incorporates unimodal subtasks to aid the learning process of the multimodal task. To enable this, we first designed Modality Recalibration Module (MRM) so that features of each modality are recalibrated based on the features of other modalities. Then, we designed a unimodal label generation module (ULGM) based on the notion that the distance between two features in a common semantic feature space is proportional to the distance between the corresponding labels in a label space. From this, ULGM was able to generate unimodal annotations from the multimodal label in a self-supervised manner, which saved a tremendous amount of human labor. The experiment results validated our notion as well as the reliability of the unimodal labels generated from our ULGM.

For future work, expanding the framework to jointly train sentiment and emotion tasks could be worthwhile. Recently (Akhtar et al., 2019) proposed that MSA and Multimodal Emotion Recognition are closely correlated; therefore their tasks can be carried out jointly. Applying contrastive learning for different emotion classes and exploiting correlation between sentiment and emotion could help achieve better results in both tasks.

Limitations

A limitation of our work is that the initial features for audio and video are extracted using off-the-shelf frameworks: COVAREP and FACET. Therefore these features are fixed and cannot be further fine-tuned unlike the text features which are fine-tuned during training. Working with fixed features, compared to dynamic features which can be adjusted via learning, inevitably results in subpar performance. We expect this limitation can be alleviated by making our framework completely end-to-end by using raw audio and video data and introducing learning-based audio and video feature extraction modules. However, using raw data can exponentially increase memory usage which is another challenge that needs to be considered. Further, by introducing additional MRM and SPT modules, our method took approximately twice the time as the (Yu et al., 2021) during inference using the unaligned MOSI dataset.[5] Double in inference time hinders the community's strive to build faster and more compact models.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This work was supported by the Institute of Information & communications Technology Planning & evaluation(IITP) grant funded by the Korea government(MSIT) (No.2020-0-00842, Development of Cloud Robot Intelligence for Continual Adaptation to User Reactions in Real Service Environments).

References

Md Shad Akhtar, Dushyant 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, Volume 1 (Long and Short Papers)*, pages 370–379, Minneapolis, Minnesota. Association for Computational Linguistics.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria,

[5] After 10 runs, the average inference time for our method was approximately 0.775 seconds, while (Yu et al., 2021) was 0.378 seconds.

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 (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

José Carlos Castillo, Álvaro Castro-González, Fernándo Alonso-Martín, Antonio Fernández-Caballero, and Miguel Ángel Salichs. 2018. Emotion detection and regulation from personal assistant robot in smart environment. In *Personal assistants: Emerging computational technologies*, pages 179–195. Springer.

Junyan Cheng, Iordanis Fostiropoulos, Barry Boehm, and Mohammad Soleymani. 2021. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2447–2458, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maartje MA De Graaf and Somaya Ben Allouch. 2013. Exploring influencing variables for the acceptance of social robots. *Robotics and autonomous systems*, 61(12):1476–1486.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

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*, pages 1122–1131.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

iMotions. 2013. Facet imotions biometric research platform.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher 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*, pages 150–161, Brussels, Belgium. Association for Computational Linguistics.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher 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 (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, Sofia, Bulgaria. Association for Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online. Association for Computational Linguistics.

Diana L Robins, Elinora Hunyadi, and Robert T Schultz. 2009. Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain and cognition*, 69(2):269–278.

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*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. 2020. Mmtm: Multimodal transfer module for cnn fusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797.

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*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

A Appendices

A.1 Dataset Split

Table 6: Train, validaton, test set split for CMU-MOSI and CMU-MOSEI datasets.

Dataset	# Train	# Valid	# Test	# All
MOSI	1284	229	686	2199
MOSEI	16326	1871	4659	22856

A.2 Hyper-parameter Settings

Table 7: Hyper-parameters used in the two datasets. The second half of the hyper-parameters (bottom row) are for the SPT.

Hyper-parameter	CMU-MOSI	CMU-MOSEI
Batch size	32	32
LR for BERT	$5e-5$	$5e-5$
LR for others	$1e-2$	$1e-3$
output dropout	0.3	0.1
# Encoder layer	4	4
# Head	8	4
Embed size	32	32
Attn dropout	0.3	0.1
ReLU dropout	0.3	0.1
Residual dropout	0.3	0.1
Embed dropout	0.3	0.2

A.3 Loss Function Ablation Study

Table 8: Loss function ablation study on the unaligned MOSI dataset. In Acc_2 and F1-Score, the left side of the "/" is the "negative/non-negative" method and the right side is the "negative/positive" method.

Loss type	F1-Score	Acc_2	MAE	Corr
L1 loss	**84.3/86.3**	**84.4/86.3**	**0.703**	**0.800**
L2 loss	80.8/81.0	80.8/81.0	0.832	0.737
Huber loss	78.1/79.2	78.2/79.2	0.818	0.744