EASUM: Enhancing Affective State Understanding through Joint Sentiment and Emotion Modeling for Multimodal Tasks

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

Abstract
Multimodal sentiment analysis (MSA) and multimodal emotion recognition (MER) tasks have gained a surge of attention in recent years. Although both tasks share common ground in many ways, they are often treated as a separate task. In this work, we propose, EASUM, a new training scheme for bridging the MSA and MER tasks. EASUM aims to bring mutual benefits to both tasks based on the premise that the sentiment and emotion are closely related; hence each information should provide deeper insight into one’s affective state to complement the other. We exploit this premise to further improve the performance of each task by 1) first training a domain general model using four benchmark datasets from the MSA and MER tasks: CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP. Depending on the dataset, the domain general model learns to predict sentiment or emotion values based on the domain invariant features. 2) Then these values are later used as auxiliary pseudo labels when training a domain specific model for each task. Our premise as well as new training scheme are validated through extensive experiments on the four benchmark datasets. The results also demonstrate that the proposed method outperforms the state-of-the-art on the CMU-MOSI, CMU-MOSEI, and MELD datasets, and performs comparable to the state-of-the-art on the IEMOCAP dataset while using approximately 40% fewer parameters.

1. Introduction
Computing one’s affective states can be a challenging, yet a niche task. Sentiment analysis and emotion recognition, tasks designed for this purpose, traditionally relied on analyzing textual data [49]. However, using text data alone can lead to suboptimal understanding of emotions or sentiment expressed in communication as humans communicate not only through words, but also through facial expressions and vocal intonation [32].

Multimodal sentiment analysis (MSA) and multimodal emotion recognition (MER) seek to address this limitation by incorporating multiple modes of communication simultaneously. Thanks to the explosion of online content in recent years, contemporary sentiment analysis and emotion recognition go beyond merely analyzing texts: these days, additional data such as tone of voice and facial expressions are also considered. This not only allows deep learning models to have a more comprehensive understanding of one’s affective states, but also helps the model to more accurately predict data that are more nuanced [39].

With the prominence of MSA and MER in the deep learning community, numerous models for these tasks have been developed [13,16,19,27,39,47,55]. However, the majority of these models target each task independently. While they are distinctive to one another, sentiment and emotion are closely related and often display high cohesiveness [34]. Sentiment is often associated with the polarity of affective states such as positive, negative, or neutral. Emotion, on the other hand, refers to a specific affective state such as happiness, sadness, anger, fear, etc. Therefore, analyzing sentiment can provide insights into the latent emotions and vice versa. For instance, a positive sentiment could come from a feeling of joy or excitement, while a negative sentiment may be driven by anger or fear. Hence, knowing both sentiment and emotion can provide a richer understanding of one’s affective state and can help each other to enhance the accuracy of both sentiment analysis and emotion recognition tasks.

Motivated by this observation, this paper aims to bring a solution to uniting MSA and MER by utilizing both sentiment and emotion information for each task. While prior works that explore this idea exist [1,17], their count remains quite minimal. We speculate this is partly due to the lack of datasets that contain both sentiment and emotion annotations, which usually stems from the labor-intensive nature of the data annotation job. To address this issue, different from the previous studies, our work approaches this problem from a domain generalization perspective. More specifically, we first train a domain general model using both MSA and MER benchmark datasets to diminish the distri-
bution gap between the these datasets so that the model can predict sentiment and emotion values based on the domain-invariant features. Subsequently, we leverage the prediction made by the domain general model to boost the performance of individual tasks.

It has been shown through many studies [4,31,36,45,52,52,53,60] that domain general models can be developed by training a model using different yet related domains. Similarly, benchmark datasets for the MSA and MER tasks are drawn from different sources, resulting in inherently distinct data distributions. Despite the difference, MSA and MER datasets share a common objective; that is, they are all used to compute individuals’ affective states. For this reason, we posit that they exhibit analogous traits within a latent semantic space. This expectation provides a foundation for implementing domain generalization in our work. To achieve this, motivated by [17], we leverage the CMU-MOSI and CMU-MOSEI datasets for the MSA task, and the MELD and IEMOCAP datasets for the MER task. We exploit the underlying similarities between these datasets by employing a domain alignment technique.

To this end, we propose EASUM, a training scheme for enhancing affective state understanding through joint sentiment and emotion modeling for multimodal tasks. EASUM is divided into two phases: 1) in the first phase, we focus on training a domain general (DG) model by aligning the four datasets at both domain and category levels. A moment matching technique is employed for domain-level alignment. To achieve category-level alignment, we harness classifiers that predict sentiment and emotion values using the domain invariant features obtained through the moment matching technique. 2) In the second phase, we focus on training a domain specific (DS) model for each task. The domain specific model has a two stream structure: the first stream employs the pretrained DG model from the first phase to generate pseudo labels; the second stream utilizes these pseudo labels as auxiliary supervision during training to enhance performance of each task.

Through this training scheme, we show that the auxiliary information gained from the DG model can indeed help boost performance for both tasks. Further, we show the quality of the pseudo labels are adequate. The main contributions of our work can be summarized as follows:

- We propose EASUM, a two phase training scheme, where in the first phase, the DG model explores the underlying commonality between the MSA and MER tasks, while in the second phase, the DS model leverages the information gained from the DG model to enhance the performance of both tasks.

- To the best of our knowledge, our work is the first work to apply domain generalization in the MSA and MER fields and build a DG model from a mix of MSA and MER benchmark datasets.

- Our training scheme is viable to other sentiment and emotion datasets and can easily be expanded to cope with more datasets.

- Our results consistently surpass the current state-of-the-art on CMU-MOSI and MELD datasets and are comparable to, and sometimes surpass the current state-of-the-art on CMU-MOSEI and IEMOCAP datasets.

2. Related Work

MSA and MER. Many MSA studies focused on improving the performance by better modeling joint representations of text, audio, and video. Some of the prime works include using multi-dimensional tensor [55], attention mechanism [56,57], multi-stage fusion [23], and Transformer architecture [6,39,47]. Recent works [19,54] have shown importance of incorporating modality-specific information in addition to the joint representation via multi-task learning. In the MER task, [18] and [56] used GCN and attention for fusion. [13] and [21] both proposed a context-aware model using a memory network and GNN to model complex dynamics and dependencies in dialogues. Moreover, there have been attempts to solve both MSA and MER tasks via multi-task learning [1] and creating universal labels [17]. However, [1] requires a dataset that contains both sentiment and emotion annotations, and [17] only uses textual information when generating the universal labels, which can lead to suboptimal understanding of one’s affective state. Our work, on the other hand, utilizes all three modalities to reach optimal understanding of one’s affective state when generating pseudo labels.

Domain Alignment. Domain alignment refers to the process of aligning feature distributions across different training domains. It is a technique that is widely used in unsupervised domain adaptation (UDA) to reduce domain shift between source and target domains. One popular method used for domain alignment is Maximum Mean Discrepancy (MMD), which reduces the distance between feature distributions of different domains [11,28,29,50]. Other commonly used methods include correlation alignment [3,37,41,42] and adversarial-based approach [10,26,48]. Further, various moment matching methods have been proposed to reduce domain discrepancy. For instance, [33] applied GAN to align the mean and covariance of two different data distributions, [36] utilized moment matching for multi-source domain adaptation, and [4] employed higher-order moment matching to better represent feature distribution in each domain. In this work, we explore both MSA and MER tasks using a moment matching method to learn domain-invariant features by minimizing distribution discrepancies.
3. Methodology

3.1. Problem Definition

Each training domain, \(D_i = \{D_1, D_2, ..., D_n\}\), has input which is composed of three types of modalities: text, audio, and video (\(X^m_i\)). The goal of the MSA and MER tasks is to take this multimodal input and predict a sentiment intensity \(\hat{y}_{s} \in \mathbb{R}\) and an emotion category \(\hat{y}_{e} \in \mathcal{Y}_i\), respectively, where \(\mathcal{Y}_i\) is an emotion label space of the \(i\)th training domain.

3.2. Model Overview

As shown in Figure 1, the proposed EASUM is divided into two phases. The first phase trains a domain general model which is composed of feature extractors (for audio and video), fusion module, multimodal representation learning module, moment matching component, and two classifiers. The moment matching component minimizes the moment-related distance to align data distributions of the four datasets. The “Senti FC” and “Emo FC” are used for classifying a sentiment polarity (positive, neutral, negative) from the MSA datasets, and an emotion category from the MER datasets using the domain invariant features, respectively. The second phase trains a domain specific model which closely resembles the first phase model excluding the moment matching component. The second phase model utilizes the domain general model to generate pseudo labels to better perform the MSA and MER tasks.

3.3. First Phase: Training DG Model

We have four training domains \(D_i \in \{1, 2, 3, 4\}\) where domains 1, 2, 3, and 4 indicate MOSI, MOSEI, MELD, and IEMOCAP, respectively. The input data and the corresponding labels for the \(i\)th training domain are \(X^m_i = \{x^m_{i,j}\}_{j=1}^{N_i}\) and \(Y_i = \{y_{i,j}\}_{j=1}^{N_i}\), where \(N_i\) is the number of \(i\)th training domain data and \(x^m_{i,j} \in \mathbb{R}^{l_m \times d_m}\), where \(l_m\) is a sequence length, and \(d_m\) is a feature dimension of \(m\)-modality. This indicates the data from different domains share the same feature space. However, not all do-
mains share the same label space. For instance, $Y_{i \in \{1, 2\}} = \{y_{i,j}\}_{j=1}^{N_i}$, where $y_{i,j} \in \mathbb{R}$ is a sentiment intensity, while $Y_{i \in \{3, 4\}} = \{y_{i,j}\}_{j=1}^{N_i}$, where $y_{i,j} \in \mathcal{Y}_i$ is an emotion category. The goal is to learn a model that is generalized to all training domains.

**Feature Extraction.** Following [17], we use acoustic data that have been processed using librosa [30] and visual data that have been extracted from efficientNet [44] and that have been pretrained on VGGFace [35] and AFEW [8] datasets.

**Data Augmentation and Processing.** We use CMU-MOSI and CMU-MOSEI datasets from the MSA task; and MELD and IEMOCAP datasets from the MER task to train a model that has general understanding of humans’ affective states. To train a DG model, we begin by balancing the number of training samples across domains by applying data augmentation. See Table 1 for the data split of each dataset. In [22], it has been found that perturbing data in the feature space with Gaussian noise during training is not only a great way to augment data, but it also leads to a classifier with domain-generalization performance. We adopt this data augmentation technique to the visual and audio features extracted from efficientNet [44] and that have been pretrained on VGGFace [35] and AFEW [8] datasets.

To train a DG model, we begin by balancing the number of training samples across domains by applying data augmentation. See Table 1 for the data split of each dataset. In [22], it has been found that perturbing data in the feature space with Gaussian noise during training is not only a great way to augment data, but it also leads to a classifier with domain-generalization performance. We adopt this data augmentation technique to the visual and audio features extracted from efficientNet [44] and that have been pretrained on VGGFace [35] and AFEW [8] datasets.

**Fusion.** The idea that word representation can shift based on acoustic and visual information was used as a way to fuse textual, acoustic, and visual data [39, 51]. Borrowing the name from [39], we also use Multimodal Adaptation Gate (MAG) to obtain a fused representation of text, audio, and video data. MAG receives text embedding, acoustic and visual features as inputs and calculates the displacement that occurs in the textual semantic space by introducing acoustic and visual data. The displacement is calculated using a gating mechanism as follows:

$$R_i = g_a(W_a \tilde{X}_i^a) + g_v(W_v \tilde{X}_i^v) + b_i,$$

with

$$g_a = ReLU(W_{ga}[\tilde{X}_i^a; \tilde{X}_i^v]) + b_a,$$

$$g_v = ReLU(W_{gv}[\tilde{X}_i^a; \tilde{X}_i^v]) + b_v,$$

where $W_{ga}$ and $W_{gv}$ are weight matrices of gating mechanism for visual and acoustic modality, and $b_a$ and $b_v$ are biases. By using this displacement, the fused representation can be computed as follows:

$$F_i = \tilde{X}_i^a + \lambda R_i,$$

with

$$\lambda = \min(\|\tilde{X}_i^a\|_2, \|\tilde{R}_i\|_2 \gamma, 1),$$

where $\gamma$ is a hyperparameter, and $\|\cdot\|_2$ is $L2$ normalization.

**Multimodal Learning.** The fused representation is used as an input to the ConVBERT encoder [20] to learn a meaningful multimodal representation. ConVBERT is an improved version of BERT [7] which uses a mixed attention block that integrates span-based dynamic convolution and self-attention. The span-based dynamic convolution can capture local dependency more effectively and efficiently by generating local relation of the input token conditioned on its local context instead of a single token. By incorporating span-based dynamic convolution head instead of relying entirely on the global self-attention block which suffers large memory footprint and computation cost, ConVBERT can better model both global and local dependencies with reduced redundancy. ConVBERT also projects the embedding feature to a smaller dimension space, adopting a bottleneck design. This significantly reduces computational costs within the self attention mechanism and forces attention heads to produce more compact and useful information. Moreover, a grouped linear operator is applied to the feed-forward to further reduce parameters while maintaining the representation power. The multimodal representation extracted from ConVBERT is denoted as follows:

$$M_i = \text{ConVBERT}(F_i; g_{\text{ConVBERT}}) \in \mathbb{R}^{N_i \times l \times d},$$

where $g_{\text{ConVBERT}}$ is the learnable parameters of ConVBERT.

**Learning Domain Invariant Features.** We use the multimodal representation obtained from ConVBERT to perform domain alignment which is a crucial aspect in training a domain general model. From numerous domain alignment techniques available, we opt for the moment matching technique, which has demonstrated its effectiveness in the field of multi-source domain adaptation [36]. To align data distribution across domains, we minimize the $k$ order moment distance between the multimodal representation of different
domains. The \( k \) order moment distance between two domains is calculated as follows:

\[
MD(D_i, D_j) = \sum_{k=1}^{2} \left\| E(M^k_i) - E(M^k_j) \right\|_2.
\]

Then the total moment distance between all training domains becomes as follows:

\[
MD_{total} = \binom{n}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} MD(D_i, D_j),
\]

where \( n \) is the number of training domains.

**DG Model Training Objective.** Lastly, the sentiment polarity (\( \hat{y}_s \)) and emotion categories (\( \hat{y}_e \)) are predicted using the multimodal representation of the MSA and MER datasets, respectively, through two independent fully connected (FC) layers as follows:

\[
\hat{y}_t = W_t M_i + b_t,
\]

where \( t \in \{s, e\} \). Note, the FC layers share parameters for the same task. The overall objective function becomes

\[
\mathcal{L}_{DG} = \sum_{i=1}^{n} \mathcal{L}_{D_i}(\hat{y}_t, y_t) + \omega MD_{total},
\]

where \( \mathcal{L}_{D_i} \) is the softmax cross entropy loss for each domain, and \( \omega \) is the trade-off parameter.

### 3.4. Second Phase: Training DS Model

After learning the DG model, we train the second phase model which has a two-stream structure. The upper stream, referred to as the DG stream, is employed to generate pseudo labels using the pretrained DG model from the first phase, where its weights remain frozen and are not further trained. The lower stream, referred to as the DS stream, is utilized for conducting the MSA and MER tasks with the assistance of pseudo labels serving as auxiliary supervision. The DS model is largely similar to the DG model except that DeBERTa [15] is used to learn a multimodal representation.

**DeBERTa [15]** is an improved version of BERT [7] and RoBERTa [61] by using a disentangled attention mechanism and enhanced mask decoder. Specifically, each word is represented with the content and position vectors, and the attention weights are calculated using disentangle matrices based on words’ contents and relative positions. DeBERTa’s new enhanced mask decoder incorporates absolute word position embedding to decode the masked words based on the aggregated contextual embeddings of word contents and relative positions. Moreover, DeBERTa uses a scale invariant fine-tuning technique to improve the training stability and generalization by applying perturbations to the normalized word embedding. Implementation of these methods allowed DeBERTa’s enhanced efficiency as well as improved performance on downstream tasks. The domain specific multimodal representation extracted from DeBERTa is denoted as

\[
M = \text{DeBERTa}(F; \theta^{DeBERTa}) \in \mathbb{R}^{B \times 1 \times d},
\]

where \( B \) is a batch size, \( F \) is the fused representation obtained from MAG, and \( \theta^{DeBERTa} \) is the learnable parameters of DeBERTa.

The multimodal representations obtained from DeBERTa are then passed to a set of two classifiers: one for predicting sentiment and the other for emotion category. One of the classifiers is trained using the annotations in the dataset, while the other classifier is trained using the pseudo labels obtained from the DG stream. For instance, in the case of the CMU-MOSI dataset which contains sentiment annotation, the DS model is trained using the sentiment annotation from the dataset as well as the emotion pseudo labels generated from the DG stream.

**Soft Pseudo Labels.** Soft labels, which are often used in semi-supervised learning [25] and knowledge distillation [53], represent the likelihood of a data sample belonging to each class; therefore, they provide a more flexible representation of the class probabilities. For this reason, soft labels can be more informative than hard labels because they provide richer information about the uncertainty of the model’s prediction [40]. Further, they reduce biases pertaining to particular datasets [9]. Motivated by this, we use soft labels for the pseudo labels generated from the DG stream, hence the name - soft pseudo label, to supervise auxiliary learning using Kullback-Leibler (KL) divergence.

**DS Model Training Objective.** Then the training objective of our DS model becomes

\[
\mathcal{L}_{DS}(\hat{y}_m, y_m, \hat{y}_a, y_a^*) = \mathcal{L}_{task}(\hat{y}_m, y_m) + \eta \mathcal{L}_{KLdiv}(\hat{y}_a, y_a^*),
\]

where \( \hat{y}_m, y_m \) are the DS model’s prediction and the ground truth label for the main task learning, and \( \hat{y}_a, y_a^* \) are the DS model’s prediction and the soft pseudo label for the auxiliary learning. \( \mathcal{L}_{task}(\hat{y}_m, y_m) \) is the MSE loss for the MSA task and softmax cross entropy loss for the MER task. \( \eta \) is the trade-off parameter, and \( \mathcal{L}_{KLdiv}(\hat{y}_a, y_a^*) \) is the KL divergence loss between the DS model’s auxiliary prediction and the soft pseudo labels, which facilitates the DS model’s auxiliary prediction to follow the soft pseudo labels.

### 4. Experimental Settings

#### 4.1. Datasets

CMU-MOSI [58] dataset contains 2,199 labeled video clips from 89 speakers. The videos are crawled from YouTube which address opinions on movies, books, and
products. Each video is annotated with sentiment on a [-3,3] range. CMU-MOSEI [59] dataset contains 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 in this work. IEMOCAP [2] dataset contains approximately 12 hours of data, including video, speech, motion capture of face, and text transcriptions. Each video is segmented into utterances which are annotated with one of six emotion labels: happy, sad, neutral, angry, excited, and frustrated as well as dimensional labels such as valence, activation and dominance. Only the emotion labels are used in this work. MELD [38] dataset contains more than 1,400 dialogues and 13,000 utterances from the Friends TV series. Each utterance is annotated with one of the seven emotion classes: anger, disgust, sadness, joy(=happy), surprise, fear, or neutral as well as sentiment polarity. In this work, we only use emotion labels from MELD.

Table 1. Data split of the four datasets and the type of annotations included in each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Train</th>
<th>Valid</th>
<th>Test</th>
<th>Senti.</th>
<th>Emo.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSI</td>
<td>1284</td>
<td>229</td>
<td>686</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>MOSEI</td>
<td>16326</td>
<td>1871</td>
<td>4659</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MELD</td>
<td>9989</td>
<td>1108</td>
<td>2610</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IEMOCAP</td>
<td>5354</td>
<td>528</td>
<td>1650</td>
<td>-</td>
<td>✓</td>
</tr>
</tbody>
</table>

4.2. Baseline Models

The baseline models for the MSA task include the following: LMF [27] performs multimodal fusion using low-rank tensors. TFN [55] models intra- and inter-modality dynamics through multi-dimensional tensors. MFM [46] factorizes representations into multimodal discriminative and modality-specific generative factors to learn multimodal data. ICCN [43] learns correlations between modalities via deep canonical correlation analysis. MulT [47] uses cross-modal attention to model interactions between asynchronous modalities and latently adapt one modality to another. MISA [14] learns modality-invariant and modality-specific features to capture a holistic view of the multimodal data. MAG-BERT [39] applies multimodal adaptation gating mechanism to BERT to model multimodal representations. MIMM [12] maximizes the mutual information in modalities and between multimodal and unimodal representations to better preserve information. Self-MM [54] generates unimodal labels for each modality and jointly trains multimodal and unimodal tasks. SUGRM [19] is an improved version of Self-MM which recalibrates each modality and maps each modality to a common latent space to facilitate unimodal label generation. UniMSE [17] generates universal labels based on the similarity in text embeddings among data samples and uses T5 model and contrastive learning to perform MSA and MER tasks.

The baseline models with which we compare our model for the MER task include: LMF, TFN, MFM, UniMSE as well as MM-DFN [16] which employs a graph-based dynamic fusion module to fuse multimodal contextual features in a conversation.

4.3. Evaluation Metrics

Following the previous works [12, 17, 19, 39, 47, 54], we evaluate our model using four metrics for the MSA task: binary F1 score (F1), binary classification accuracy (Acc2), Mean Absolute Error (MAE), and Pearson correlation (Corr). For the MER task, we evaluate our model using two metrics: accuracy (Acc) and weighted F1 score (w-F1).

4.4. Implementation Details

We trained our framework using NVIDIA TITAN Xp and Intel i7-9700K. We use the batch size of 48 and AdamW as the optimizer. We set the learning rate to $3.5e-5$ for the IEMOCAP dataset and $1e-5$ for the rest of the datasets. We use 8 and 3 DeBERTa encoder layers when training the DS model for the MSA and MER task, respectively, and use 8 ConvBERT encoder layers for the DG model. The feature dimension of the acoustic and visual representations is 64, and the embedding size for both ConvBERT and DeBERTa is 768. The sequence lengths of the text, acoustic, visual representations are 40, 157, 32, respectively. We set $\gamma$ to 1, $\omega$ to 0.1, and $\eta$ to 0.5 for the MOSI and MOSEI datasets, 0.1 for the IEMOCAP dataset, and 1 for the MELD dataset.

5. Results and Analysis

5.1. Quantitative Results

Table 2 shows the experimental results for the MSA task on both CMU-MOSI and CMU-MOSEI datasets. As can be seen in the table, our model set the new SOTA record on both CMU-MOSI and CMU-MOSEI datasets. For the CMU-MOSEI dataset, particularly achieving a notable improvement on the Corr metric. Further, our model outperformed the previous SOTA results on the MAE and Corr metrics. For the CMU-MOSI dataset, in spite of our model’s shortcoming compared to the previous SOTA results on the F1 and Acc2 metrics, the performance gap between our model and the previous SOTA model [17] is minuscule (only short by 0.16% and 0.1% on F1 and Acc2 metrics). Additionally, Table 3 shows the experimental results for the MER task on both MELD and IEMOCAP datasets. We only compare our model with the prior multimodal models.
Table 2. Experimental results of our model compared to the baseline models on the CMU-MOSI and CMU-MOSEI datasets for the MSA task. The bold numbers indicate the best performance, and ↑ indicates higher number is better, while ↓ indicates lower number is better.

<table>
<thead>
<tr>
<th>Model</th>
<th>MOSI</th>
<th></th>
<th>MOSEI</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1(%) ↑</td>
<td>Acc2(%) ↑</td>
<td>MAE ↓</td>
<td>Corr ↑</td>
</tr>
<tr>
<td>LMF [27]</td>
<td>82.4</td>
<td>82.5</td>
<td>0.917</td>
<td>0.695</td>
</tr>
<tr>
<td>TFN [55]</td>
<td>80.7</td>
<td>80.8</td>
<td>0.901</td>
<td>0.698</td>
</tr>
<tr>
<td>MFM [46]</td>
<td>81.6</td>
<td>81.7</td>
<td>0.877</td>
<td>0.706</td>
</tr>
<tr>
<td>ICCN [43]</td>
<td>83.0</td>
<td>83.0</td>
<td>0.862</td>
<td>0.714</td>
</tr>
<tr>
<td>MulT [47]</td>
<td>83.9</td>
<td>84.1</td>
<td>0.861</td>
<td>0.711</td>
</tr>
<tr>
<td>MISA [14]</td>
<td>82.0</td>
<td>82.1</td>
<td>0.804</td>
<td>0.764</td>
</tr>
<tr>
<td>MAG-BERT [39]</td>
<td>86.0</td>
<td>86.1</td>
<td>0.712</td>
<td>0.796</td>
</tr>
<tr>
<td>MIMM [12]</td>
<td>86.0</td>
<td>86.1</td>
<td>0.700</td>
<td>0.800</td>
</tr>
<tr>
<td>Self-MM [54]</td>
<td>86.0</td>
<td>86.0</td>
<td>0.713</td>
<td>0.798</td>
</tr>
<tr>
<td>SUGRM [19]</td>
<td>86.3</td>
<td>86.3</td>
<td>0.703</td>
<td>0.800</td>
</tr>
<tr>
<td>UniMSE [17]</td>
<td>86.42</td>
<td>86.90</td>
<td>0.691</td>
<td>0.809</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>87.18</strong></td>
<td><strong>87.20</strong></td>
<td><strong>0.663</strong></td>
<td><strong>0.828</strong></td>
</tr>
</tbody>
</table>

Table 3. Experimental results of our model compared to the baseline models on the MELD and IEMOCAP datasets for the MER task. The bold numbers indicate the best performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>MELD</th>
<th>IEMOCAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w-F1 ↑</td>
<td>Acc ↑</td>
</tr>
<tr>
<td>LMF [27]</td>
<td>58.30</td>
<td>61.15</td>
</tr>
<tr>
<td>TFN [55]</td>
<td>57.74</td>
<td>60.70</td>
</tr>
<tr>
<td>MFM [46]</td>
<td>57.80</td>
<td>60.80</td>
</tr>
<tr>
<td>MM-DFN [16]</td>
<td>59.46</td>
<td>62.49</td>
</tr>
<tr>
<td>UniMSE [17]</td>
<td>65.51</td>
<td>65.09</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>65.93</strong></td>
<td><strong>66.70</strong></td>
</tr>
</tbody>
</table>

Table 4. An ablation study on introducing additional modalities on the CMU-MOSI dataset. The bold numbers indicate the best performance, and the underlined numbers indicate the enhanced performance by incorporating V or A modality to the T modality.

<table>
<thead>
<tr>
<th>Modality</th>
<th>F1(%) ↑</th>
<th>Acc2(%) ↑</th>
<th>MAE ↓</th>
<th>Corr ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>86.00</td>
<td>85.98</td>
<td>0.789</td>
<td>0.772</td>
</tr>
<tr>
<td>T, A</td>
<td>85.95</td>
<td>85.97</td>
<td>0.746</td>
<td>0.793</td>
</tr>
<tr>
<td>T, V</td>
<td>86.44</td>
<td>86.43</td>
<td>0.741</td>
<td>0.797</td>
</tr>
<tr>
<td>T, A, V</td>
<td><strong>87.18</strong></td>
<td><strong>87.20</strong></td>
<td><strong>0.663</strong></td>
<td><strong>0.828</strong></td>
</tr>
</tbody>
</table>

5.2. Ablation Study

We show the results of an ablation study exploring how introducing additional modalities contributes to the model’s performance using CMU-MOSI dataset in Table 4. The table shows that text is a dominant modality, showing sufficient performance alone, which is in line with the previous finding [47]. The table also shows the importance of visual and acoustic information. As can be seen in the table, the performance of the model generally increases as we incorporate more modalities with the exception of the combination of text and audio modality (T, A) on the F1 and Acc2 metrics. However, from the MAE and Corr perspective, the model performance grows with the addition of modalities. This demonstrates that combining text with visual and/or acoustic information can capture more nuanced affective states that the text alone cannot otherwise. Further, the model exhibited its optimal performance when all modalities were used. This observation indicates that the model is most capable of capturing an individual’s affective state comprehensively when all three modalities are utilized. Moreover, we can infer from the results that the visual modality boosts the model’s performance more than the acoustic modality.

In addition, in order to investigate the efficacy of the auxiliary pseudo labels, we compare the model’s performance with and without the soft pseudo labels (SPL) in Table 5. As can be seen in the last row of the table, including soft pseudo labels consistently reinforces the model’s performance gain. This empirically demonstrates the effectiveness of introducing auxiliary information via soft pseudo labels in the MSA and MER tasks. To further elaborate, in the case of the MSA tasks, it notably benefited from introducing the auxiliary emotion pseudo labels. However, the benefit the sentiment pseudo labels brought to the MER tasks was...
In this paper, we introduced a new training scheme named EASUM for the MSA and MER tasks, which aims to enhance performance of both tasks by leveraging the interrelation between sentiment and emotion. Specifically, our approach is predicated on the idea that knowing both information (sentiment and emotion) can offer a more profound understanding of an individual’s affective state than knowing just one information. To explore this idea, we utilized four benchmark datasets from the MSA and MER tasks and trained the domain general model to bridge the gap among the four domains. Then, we used the domain general model to produce pseudo labels to serve as additional guidance when training the domain specific model for each task. We investigated the impact of the pseudo labels on the performance of each task and validated the effectiveness of our training scheme through the experiments. Further, we showed the adequacy and reliability of the pseudo labels generated from the domain general model. Through this training scheme, our model was able to achieve new SOTA results on the CMU-MOSEI (on MAE, Corr metrics), CMU-MOSI, and MELD datasets, as well as achieve nearly SOTA results on the IEMOCAP dataset while using approximately 40% fewer parameters compared to the previous SOTA model, all without requiring labor-intensive data annotation job for the additional auxiliary labels.

**Acknowledgement.** 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).

---

1For the IEMOCAP dataset, only the visual and acoustic features are available to the public (no raw video or audio available). Therefore, we omit the acoustic and visual part of the IEMOCAP dataset in Table 6.
References


[15] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*, 2021. 5


