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 distribution 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

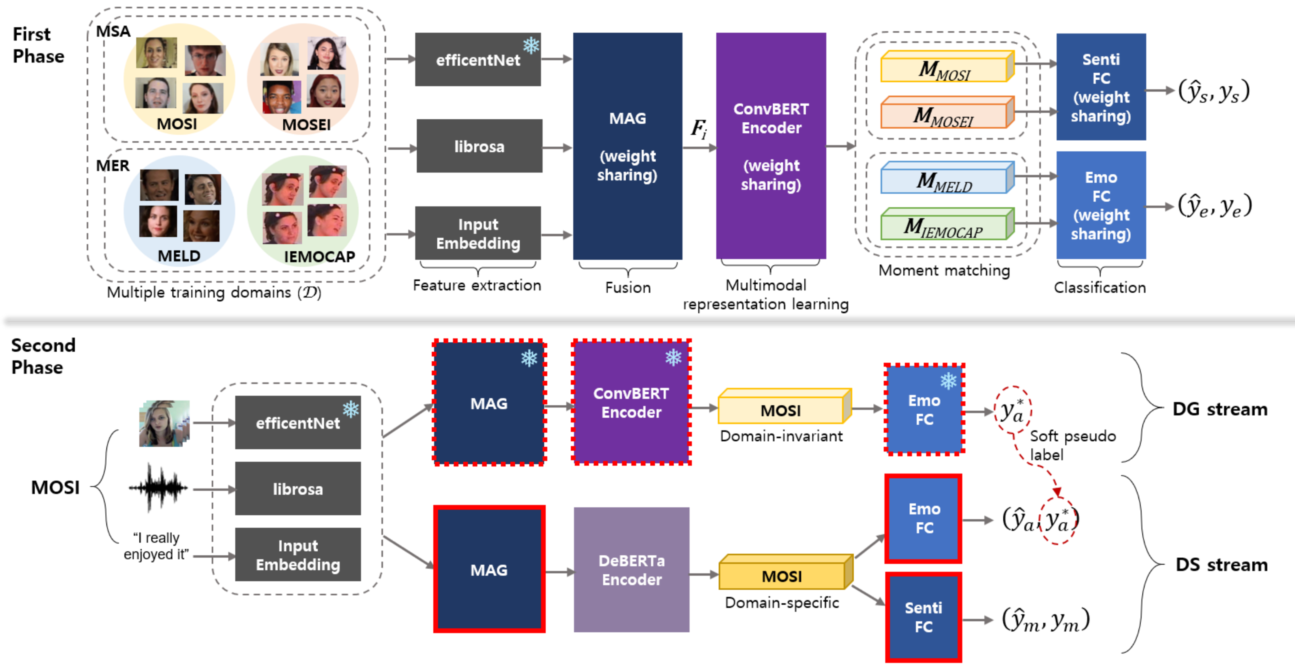


Figure 1. The overall architecture of EASUM. The upper half of the figure shows the first phase of our model whose purpose is to train a domain general (DG) model which can classify a sentiment polarity ($\hat{y}_s$) from the CMU-MOSI, -MOSEI datasets and an emotion category ($\hat{y}_e$) from the MELD, IEMOCAP datasets based on the domain-invariant features. The lower half of the model shows the second phase of our model whose purpose is to perform the MSA and MER tasks using the pretrained DG model to obtain pseudo labels (y_a^*) for auxiliary supervised learning. For example, in case of the CMU-MOSI dataset which only contains sentiment annotations (y_m), the pretrained DG model is used to generate pseudo labels for an emotion category (y_a^*) for the auxiliary supervised learning. The red dashed box indicates the pretrained weights from the first phase are loaded in the second phase without further updates (❄ indicates the model is frozen), while the red solid box indicates the pretrained weights from the first phase are loaded and further updated through training.

across the MSA and MER datasets. Subsequently, we address each task independently while leveraging the information yielded from the domain-invariant features to enhance the performance of each task.

3. Methodology

3.1. Problem Definition

Each training domain, $\mathcal{D}_i = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$, has input which is composed of three types of modalities-text, audio, and video ($\boldsymbol{X}_i^m$), where $m \in \{t, a, v\}$. 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 $\mathcal{D}_{i\in\{1,2,3,4\}}$ 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 $\boldsymbol{X}_i^m = \{\boldsymbol{x}_{i,j}^m\}_{j=1}^{N_i}$ and $Y_i = \{y_{i,j}\}_{j=1}^{N_i}$, where N_i is the number of i^{th} training domain data and $\boldsymbol{x}_{i,j}^m \in \mathbb{R}^{l_m \times d_m}$, 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 domains 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] that has 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 librosa [30] as follows:

$$\hat{\boldsymbol{X}}_i^m = \alpha \odot \boldsymbol{X}_i^m + \beta,$$

where $\alpha \in \mathbb{R}^{N_i \times l_m \times d_m}$ and $\beta \in \mathbb{R}^{N_i \times l_m \times d_m}$ are the noise samples taken from normal distribution, and $\odot$ denotes element-wise multiplication. Specifically, $\alpha \sim \mathcal{N}(1, \boldsymbol{I})$ and $\beta \sim \mathcal{N}(1, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix. For the text data augmentation, we perform a random temporal zero-out [5, 24], an efficient textual data augmentation method which randomly drops parts of the information within an input sentence, after the input embedding.

Note that the emotion label spaces for the MELD and IEMOCAP datasets are incongruent. In order to create a new comprehensive label space for the MER task, we merge emotion categories in each dataset. The resulting common emotion label space, $\mathcal{Y}$, consists of 9 emotion categories: happy, angry, sad, neutral, excited, frustrated, fear, disgust, and surprise. We also convert the sentiment intensity ($y_s \in \mathbb{R}$) to the sentiment polarity ($y_s \in \mathbb{R}^3$) in the MOSI and MOSEI 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:

$$\boldsymbol{R}_i = \boldsymbol{g}_a(\boldsymbol{W}_a \hat{\boldsymbol{X}}_i^a) + \boldsymbol{g}_v(\boldsymbol{W}_v \hat{\boldsymbol{X}}_i^v) + \boldsymbol{b},$$

with

$$\boldsymbol{g}_a = ReLU(\boldsymbol{W}_{ga}[\hat{\boldsymbol{X}}_i^t; \hat{\boldsymbol{X}}_i^a]) + \boldsymbol{b}_a,$$

$$\boldsymbol{g}_v = ReLU(\boldsymbol{W}_{gv}[\hat{\boldsymbol{X}}_i^t; \hat{\boldsymbol{X}}_i^v]) + \boldsymbol{b}_v,$$

where $\boldsymbol{W}_{ga}$ and $\boldsymbol{W}_{gv}$ are weight matrices of gating mechanism for visual and acoustic modality, and $\boldsymbol{b}_a$ and $\boldsymbol{b}_v$ are biases. By using this displacement, the fused representation can be computed as follows:

$$\boldsymbol{F}_i = \hat{\boldsymbol{X}}_i^t + \lambda \boldsymbol{R}_i,$$

with

$$\lambda = min(\frac{\|\hat{\boldsymbol{X}}_i^t\|_2}{\|\boldsymbol{R}_i\|_2}\gamma, 1),$$

where γ 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:

$$\boldsymbol{M}_i = ConvBERT(\boldsymbol{F}_i; \theta^{ConvBERT}) \in \mathbb{R}^{N_i \times l \times d},$$

where $\theta^{ConvBERT}$ 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:

$$MD(\mathcal{D}_i, \mathcal{D}_j) = \sum_{k=1}^{2} \|\mathbb{E}(\boldsymbol{M}_i^k) - \mathbb{E}(\boldsymbol{M}_j^k)\|_2.$$

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

$$MD_{total} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} MD(\mathcal{D}_i, \mathcal{D}_j),$$

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:

$$\hat{y}_t = \boldsymbol{W}_t \boldsymbol{M}_i + b_t,$$

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

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

where $\mathcal{L}_{\mathcal{D}_i}$ is the softmax cross entropy loss for each domain, and ω 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

$$\boldsymbol{M} = DeBERTa(\boldsymbol{F}; \theta^{DeBERTa}) \in \mathbb{R}^{B \times l \times d},$$

where B is a batch size, $\boldsymbol{F}$ 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]. Futher, 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. η 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.

Dataset	Train	Valid	Test	Senti.	Emo.
MOSI	1284	229	686	✓	-
MOSEI	16326	1871	4659	✓	✓
MELD	9989	1108	2610	✓	✓
IEMOCAP	5354	528	1650	-	✓

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 (Acc_2), 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 γ to 1, ω to 0.1, and η 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 the CMU-MOSI dataset across all evaluation metrics and either surpassed or achieved nearly the SOTA results on the CMU-MOSEI dataset. For the CMU-MOSEI dataset, in spite of our model's shortcoming compared to the previous SOTA results on the F1 and Acc_2 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 Acc_2 metrics). Further, our model outperformed the previous SOTA results on the MAE and Corr metrics for the CMU-MOSEI dataset, particularly achieving a notable improvement on the Corr metric.

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.

Model	MOSI				MOSEI			
	F1(%) ↑	Acc_2(%) ↑	MAE ↓	Corr ↑	F1(%) ↑	Acc_2(%) ↑	MAE ↓	Corr ↑
LMF [27]	82.4	82.5	0.917	0.695	82.1	82.0	0.623	0.700
TFN [55]	80.7	80.8	0.901	0.698	82.1	82.5	0.593	0.677
MFM [46]	81.6	81.7	0.877	0.706	84.3	84.4	0.568	0.703
ICCN [43]	83.0	83.0	0.862	0.714	84.2	84.2	0.565	0.704
MulT [47]	83.9	84.1	0.861	0.711	82.3	82.5	0.580	0.703
MISA [14]	82.0	82.1	0.804	0.764	84.0	84.2	0.568	0.717
MAG-BERT [39]	86.0	86.1	0.712	0.796	84.7	84.8	0.567	0.742
MIMM [12]	86.0	86.1	0.700	0.800	85.9	86.0	0.526	0.772
Self-MM [54]	86.0	86.0	0.713	0.798	85.3	85.2	0.530	0.765
SUGRM [19]	86.3	86.3	0.703	0.800	85.1	85.0	0.541	0.758
UniMSE [17]	86.42	86.90	0.691	0.809	**87.46**	**87.50**	0.523	0.773
Ours	**87.18**	**87.20**	**0.663**	**0.828**	87.30	87.40	**0.520**	**0.791**

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.

Model	MELD		IEMOCAP	
	w-F1 ↑	Acc ↑	w-F1 ↑	Acc ↑
LMF [27]	58.30	61.15	56.49	56.50
TFN [55]	57.74	60.70	55.13	55.02
MFM [46]	57.80	60.80	61.60	61.24
MM-DFN [16]	59.46	62.49	68.18	68.21
UniMSE [17]	65.51	65.09	**70.66**	**70.56**
Ours	**65.93**	**66.70**	69.75	70.10

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.

Modality	F1(%) ↑	Acc_2(%) ↑	MAE ↓	Corr ↑
T	86.00	85.98	0.789	0.772
T, A	85.95	85.97	<u>0.746</u>	<u>0.793</u>
T, V	<u>86.44</u>	<u>86.43</u>	<u>0.741</u>	<u>0.797</u>
T, A, V	**87.18**	**87.20**	**0.663**	**0.828**

and omit models which only use textual data for a fair comparison. Our model surpassed the previous SOTA results on all metrics on the MELD datasets and achieved comparable results on the IEMOCAP dataset. It is worth noting that our model achieved above or close to the SOTA results in the MER tasks while using only 62% of the parameters used in the previous SOTA model [17], which used T5-base as their base model (approx. 222.9M trainable parameters for the T5 base model alone vs. approx. 138.5M trainable parameters for our entire DS model).

5.2. Ablation Study

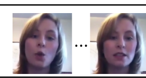
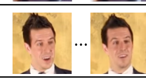
We show the results of an ablation study exploring how introducing additional modalities contributes to the model's performance using CMU-MOSI dataset in Table 4. The table shows that text is a dominant modality, showing sufficient performance alone, which is in line with the previous finding [47]. The 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 Acc_2 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

Table 5. An ablation study on the contribution of the soft pseudo labels (SPL). "w/o SPL" indicates the model's performance without the soft pseudo labels, and "w/ SPL" indicates the model's performance with the soft pseudo labels. "Improv." quantitatively shows the performance gain by introducing the soft pseudo labels.

Model	MOSI				MOSEI				MELD		IEMOCAP	
	F1 ↑	Acc_2 ↑	MAE ↓	Corr ↓	F1 ↑	Acc_2 ↑	MAE ↓	Corr ↓	w-F1 ↑	Acc ↑	w-F1 ↑	Acc ↑
w/o SPL	86.17	86.13	0.763	0.807	86.50	86.50	0.544	0.787	65.32	66.32	69.10	69.87
w/ SPL	87.18	87.20	0.663	0.828	87.30	87.40	0.520	0.791	65.93	66.70	69.75	70.10
Improv.	**1.01** ↑	**1.07** ↑	**0.1** ↓	**0.021** ↑	**0.8** ↑	**0.9** ↑	**0.024** ↓	**0.004** ↑	**0.61** ↑	**0.38** ↑	**0.65** ↑	**0.23** ↑

Table 6. Four samples from each of the four datasets. "Annot." indicates the sample's annotation from the dataset, and SPL indicates the soft pseudo labels generated from the DG stream. We show the top 3 (out of 9) emotion soft pseudo labels for the MOSI and MOSEI datasets, where "neu", "hap", and "exc" denote neutral, happy, and excited.

Dataset	Text	Acoustic	Visual	Annot.	SPL
MOSI	"Rango kind of falls into that."	slow paced and calm		0.0	neu: ≈ 0.75, sad: ≈ 0.08, hap: ≈ 0.04
MOSEI	"It's only getting better from now."	fast paced and high-spirited		2.3	hap: ≈ 0.42, neu: ≈ 0.39, exc: ≈ 0.09
MELD	"You're a genius!"	low pitched and dramatic		surprise	neu: ≈ 0.02, pos: ≈ 0.98, neg: ≈ 0.00
IEMOCAP	"I'm gonna forget him."	-	-	disgust	neu: ≈ 0.00, pos: ≈ 0.00, neg: ≈ 0.99

marginal compared to the opposite case. We suspect this is because additional emotion value provides a more specific angle to one's sentiment, while additional sentiment value can be vague. For instance, fear clearly indicates negative sentiment, while negative sentiment cannot clearly indicate whether one is disgusted or feared.

5.3. Qualitative Results

To evaluate the quality of the soft pseudo labels generated from the DG stream, we display four samples from each of the four datasets in Table 6[1]. We observe that the generated soft pseudo labels are generally in parallel with the annotations from the datasets. This shows that the soft pseudo labels generated from the DG stream are able to properly provide complementary information to the model, contributing to the model's enhanced performance. This further confirms the efficacy and the objective of the soft pseudo labels.

6. Conclusion

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).

[1]For 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

[1] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 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, June 2019. Association for Computational Linguistics. 1, 2

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359, 2008. 6

[3] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3296–3303, 2019. 2

[4] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3422–3429, 2020. 2

[5] Jiaao Chen, Dinghan Shen, Weizhu Chen, and Diyi Yang. HiddenCut: Simple data augmentation for natural language understanding with better generalizability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4380–4390, Online, Aug. 2021. Association for Computational Linguistics. 4

[6] Junyan Cheng, Iordanis Fostiropoulos, Barry Boehm, and Mohammad Soleymani. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2447–2458, 2021. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, 2019. 4, 5

[8] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012. 4

[9] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 2

[11] Muhammad Ghifary, W Bastiaan Kleijn, and Mengjie Zhang. Domain adaptive neural networks for object recognition. In *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*, pages 898–904. Springer, 2014. 2

[12] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 6, 7

[13] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604, 2018. 1, 2

[14] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery. 6, 7

[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

[16] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE, 2022. 1, 6, 7

[17] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 1, 2, 4, 6, 7

[18] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675, Online, Aug. 2021. Association for Computational Linguistics. 2

[19] Yewon Hwang and Jong-Hwan Kim. Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 35–46, 2023. 1, 2, 6, 7

[20] Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Convbert: Improving bert with span-based dynamic convolution. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, vol-

ume 33, pages 12837–12848. Curran Associates, Inc., 2020. 4

[21] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States, July 2022. Association for Computational Linguistics. 2

[22] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M. Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8886–8895, October 2021. 4

[23] Paul Pu Liang, Ziyin Liu, Amir Zadeh, and Louis-Philippe Morency. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920*, 2018. 2

[24] Ronghao Lin and Haifeng Hu. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 511–523, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 4

[25] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, and Fu-Sheng Gou. Semi-supervised linear discriminant clustering. *IEEE Transactions on Cybernetics*, 44(7):989–1000, 2014. 5

[26] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2

[27] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018. 1, 6, 7

[28] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 2

[30] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015. 4

[31] Eduardo Fernandes Montesuma and Fred Maurice Ngole Mboula. Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16785–16793, 2021. 2

[32] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, 2011. 1

[33] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. Mcgan: Mean and covariance feature matching gan. In *International conference on machine learning*, pages 2527–2535. PMLR, 2017. 2

[34] Pansy Nandwani and Rupali Verma. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81, 2021. 1

[35] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 4

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2, 4

[37] Xingchao Peng and Kate Saenko. Synthetic to real adaptation with generative correlation alignment networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1982–1991. IEEE, 2018. 2

[38] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. 6

[39] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020. 1, 2, 4, 6, 7

[40] Kusha Sridhar, Wei-Cheng Lin, and Carlos Busso. Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2021. 5

[41] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016. 2

[42] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 2

[43] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8992–8999, Apr. 2020. 6, 7

[44] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 4

[45] Mamatha Thota and Georgios Leontidis. Contrastive domain adaptation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2209–2218, 2021. 2

[46] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *ICLR*, 2019. 6, 7

[47] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 1, 2, 6, 7

[48] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[49] G Vinodhini and RM Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012. 1

[50] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 402–410, 2018. 2

[51] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019. 4

[52] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16643–16653, 2021. 2

[53] Tongkun Xu, Weihua Chen, Pichao WANG, Fan Wang, Hao Li, and Rong Jin. CDTrans: Cross-domain transformer for unsupervised domain adaptation. In *International Conference on Learning Representations*, 2022. 2, 5

[54] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797, 2021. 2, 6, 7

[55] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 1, 2, 6, 7

[56] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[57] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[58] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 5

[59] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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, 2018. 6

[60] Sicheng Zhao, Guangzhi Wang, Shanghang Zhang, Yang Gu, Yaxian Li, Zhichao Song, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. Multi-source distilling domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12975–12983, 2020. 2

[61] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China, Aug. 2021. Chinese Information Processing Society of China. 5