ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis

Jiuding Yang [*1], **Yakun Yu** [*1], **Di Niu** [1], **Weidong Guo** [†2], **Yu Xu** [2]
[1]University of Alberta
[2]Platform and Content Group, Tencent
[1]{jiuding,yakun2,dniu}@ualberta.ca
[2]{weidongguo,henrysxu}@tencent.com

Abstract

Multimodal Sentiment Analysis aims to predict the sentiment of video content. Recent research suggests that multimodal sentiment analysis critically depends on learning a good representation of multimodal information, which should contain both modality-invariant representations that are consistent across modalities as well as modality-specific representations. In this paper, we propose ConFEDE, a unified learning framework that jointly performs contrastive representation learning and contrastive feature decomposition to enhance representation of multimodal information. It decomposes each of the three modalities of a video sample, including text, video frames, and audio, into a similarity feature and a dissimilarity feature, which are learned by a contrastive relation centered around text. We conducted extensive experiments on CH-SIMS, MOSI and MOSEI to evaluate various state-of-the-art multimodal sentiment analysis methods. Experimental results show that ConFEDE outperforms all baselines on these datasets on a range of metrics.

1 Introduction

Multimodal deep learning involves interpreting and analyzing multimodal signals together, where each modality refers to a way in which something is experienced and felt, e.g., the visual, audio, or language modality. With the widespread popularity of online social media, such as Instagram, TikTok, Facebook, etc., videos containing multiple modalities have become a major information carrier, which brings new challenges to content recommendation and classification, e.g., video question answering (Lei et al., 2021; Li et al., 2020), video captioning (Ging et al., 2020; Li et al., 2020), and video retrieval (Akbari et al., 2021; Lei et al., 2021).

While traditional sentiment analysis is mainly based on language, multimodal sentiment analysis (MSA) predicts the human emotion by utilizing extra information available in visual and audio modalities of the content to assist with language-based prediction. Here, the text modality contains the semantic meaning of the spoken language. The visual modality extracts the facial characteristics (e.g., head orientation, facial expressions, and pose) of the speaker. The audio modality reflects the emphasis on the utterance (e.g., through pitch, bandwidth and intensity). MSA has recently gained much attention in research for several reasons. On one hand, because of the abundance of social media content, commercial interests are switching from gauging user opinions/emotions from text only to more thorough multimodal analysis based on videos. On the other hand, short video platforms (e.g., TikTok, Instagram) allow users to easily create multimodal content including visual information, audio, and inserted text, while these modalities are sometimes noisy or even contradicting each other in sentiments. Therefore, the presence of multimodal information in addition to the text or language itself is necessary to make a thorough conclusion about the overall sentiment of a video.

Multimodal fusion has become essential to gaining a deeper understanding of these video scenes (Baltrušaitis et al., 2018) and has proven to be helpful in many downstream tasks. Various multimodal fusion techniques have been proposed for MSA, among which a basic solution is concatenating the extracted feature of each modality before performing downstream regression or classification. Recent work has recognized the importance of identifying modality-invariant information across modalities and fuse them to strengthen sentiment prediction (Hazarika et al., 2020; Zadeh et al., 2018a; Rahman et al., 2020; Sun et al., 2020).

Although modality-invariant information helps reinforce the understanding of the content, there are also cases where sentiments of different modalities contradict each other. For example, when one

[*]These authors contributed equally to this work.
[†]Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7617–7630

thanks someone with phrases like "Finally I can rest easy tonight" or "I can't thank you enough", it is very hard to conclude whether the sentiment is positive or negative without looking at the non-verbal cues, such as tones, facial expressions, and gestures. In fact, many sarcastic opinions are expressed by non-linguistic markers. In these cases, the overall sentiment cannot simply be judged by a majority vote among all modalities. Thus, multimodal representation learning that respects both consistency and incongruity between modalities have recently shown great promise (Yu et al., 2020; Hazarika et al., 2020).

In this paper, we propose **ConFEDE**, a **Con**trastive **FE**ature **DE**composition framework, which integrates both modality decomposition within each sample and supervised contrastive learning across samples in a single unified contrastive learning framework. Our main contributions are summarized as follows: (1) We integrate inter-sample contrastive learning and intra-sample modality decomposition into a simple unified loss function, based on a customized data sampler that allows us to sample positive/negative data pairs to perform both learning tasks. (2) We propose to decompose each modality into a *similarity feature* and a *dissimilarity feature*, and use the similarity feature of the text as an anchor to build the contrastive relation among all decomposed features. This is due to the observation that sentiment analysis is still largely centered around text and spoken language, while other modalities can provide extra information to assist with prediction. (3) Based on multimodal representation learning proposed above, we further introduce a multi-task prediction loss that depends on each decomposed modality representation and enables the model to learn from both multimodal prediction and unimodal prediction.

We mainly evaluated ConFEDE on CH-SIMS (Yu et al., 2020) benchmark, which contains both unimodal and overall sentiment labels for each sample. The result shows that the proposed method significantly outperforms a wide range of state-of-the-art multimodal sentiment analysis methods. To test the capability when no unimodal labels are provided, we further conduct experiments on MOSI (Zadeh et al., 2018a) and MOSEI (Zadeh et al., 2018b), which contain only an overall sentiment label for each sample, which shows that our proposed method can also achieve better performance than state-of-the-art methods on a number of performance metrics without unimodal labels. We provide extensive ablation studies to show the effectiveness and necessity of each design component in ConFEDE. The code is released at https://github.com/XpastaX/ConFEDE/.

2 Related Work

In this section, we discuss the related work in MSA and contrastive representation learning.

2.1 Multimodal Sentiment Analysis

Prior works on multimodal sentiment analysis mostly focus on predicting sentiments based on text and vision (Zhu et al., 2022; Ji et al., 2019; Liu et al., 2019).However, there is growing interest in analyzing sentiment using all three modalities: text, audio, and vision (Yu et al., 2020, 2021; Rahman et al., 2020). Zadeh et al. (2016) were among the first to propose a multimodal dictionary that could learn the dynamic interactions between facial gestures and spoken words to model sentiments. They later introduced a Tensor Fusion Network (TFN) to learn the intra-modality and inter-modality dynamics of three modalities in an end-to-end way (Zadeh et al., 2017). Furthermore, they presented a Memory Fusion Network (MFN) which is composed of Long Short Term Memories (LSTMs) to learn the view-specific and cross-view interactions of three views (text, video, and audio) to improve sentiment analysis performance. Rahman et al. (2020) proposed a Multimodal Adaptation Gate (MAG) to fine-tune BERT (Devlin et al., 2019) on multimodal data to improve sentiment analysis performance. However, these prior works do not consider modality-specific information.

To better study the impacts that modality-specific information can bring to MSA, Yu et al. (2020) construct a new multimodal sentiment analysis dataset CH-SIMS, which contains a unimodal label for each modality of a sample. Experiments show a great improvement in overall sentiment prediction after simply integrating unimodal predictions as subtasks in the learning objective.

Hazarika et al. (2020) further decompose each modality into a modality-invariant and a modality-specific representation, and employ squared Frobenius norm loss as the regularizer. However, they treat all modalities equally while regularizing the prediction result, which ignores the different effectiveness of modalities. In real cases, the text is

usually more effective on MSA tasks compared to vision and audio. In other words, it is less "noisy" than the other two modalities. Also, they employ Central Moment Discrepancy loss to push the modality-invariant representations close and a Frobenius norm to push modality-specific representations to be orthogonal, while in our method, we integrate the above mechanism into a single loss function. Moreover, they regularize the decomposed features by reconstructing the original features with the generated features. We, instead, avoid using such a method and regularize the decomposed features with unimodal prediction tasks. To improve the decomposition performance, we further aggregate the supervised contrastive learning between samples into our frameworks by a custom-designed sampling method.

A concurrent work HyCon (Mai et al., 2021) introduces a contrastive learning method for MSA, taking both inter-sample and intra-sample contrasts into consideration. However, they ignore the regularization for each decomposed feature. In contrast, in ConFEDE, within-sample feature contrasts are constructed based on a specific pattern centered around text similarity features. Also, when performing inter-sample contrastive learning, HyCon samples positive and negative pairs randomly based on MSA labels. In contrast, we design a data sampler that considers both the labels and similarities between modalities to retrieve positive/negative pairs. Due to these reasons, our method beats HyCon on most metrics on MOSI (Zadeh et al., 2018a) and MOSEI (Zadeh et al., 2018b), and is able to utilize unimodal labels to further boost performance, e.g., on CH-SIMS (Yu et al., 2020).

2.2 Contrastive Representation Learning

Contrastive learning has achieved great success in representation learning by contrasting positive pairs against negative pairs (Akbari et al., 2021; Hassani and Khasahmadi, 2020; Chen et al., 2020). Akbari et al. (2021) train a Video-Audio-Text Transformer (VATT) using multimodal contrastive learning for the alignment of video-text and video-audio pairs, and thus achieve state-of-the-art on various computer vision tasks (e.g., audio classification and video action recognition). Hassani and Khasahmadi (2020) propose to learn node and graph level representations by contrasting encodings obtained from different structural views of graphs and achieve the state-of-the-art on various graph classification benchmarks. Chen et al. (2020) present a self-supervised framework, *SimCLR*, to learn visual representations through a contrastive loss between augmented views of the same image sample.

Khosla et al. (2020) extend self-supervised contrastive learning to the supervised setting, *i.e.*, contrasting samples from different classes. They also claim that the supervised setting is more stable for hyperparameters. We design a novel contrastive learning framework that utilizes the contrasts of modalities both within a sample and between samples to enhance multimodal representation in a unified contrastive loss guided by a specific pairing pattern. Furthermore, we propose a data sampler to retrieve similar samples as positive pairs, which is in contrast to the above prior work that obtains positive pairs by data augmentation.

3 Methodology

In this section, we introduce the overall architecture of ConFEDE followed by a detailed description of the contrastive feature decomposition process for learning multimodal representations.

3.1 Model Architecture

The overall architecture of ConFEDE is shown in Figure 1. Given a sample, we first encode each modality with corresponding feature extractors. Specifically, we use the [CLS] tag of BERT to encode text (i.e., $\mathbf{T}$), and two separate transformer encoders to encode vision and audio modalities (i.e., $\mathbf{V}$ and $\mathbf{A}$), respectively. After that, we decompose each encoded modality into a similarity feature (i.e., $\mathbf{T}_s$/$\mathbf{V}_s$/$\mathbf{A}_s$ in Figure 1) and a dissimilarity feature (i.e., $\mathbf{T}_d$/ $\mathbf{V}_d$/$\mathbf{A}_d$ in Figure 1) with different projectors. Each projector is composed of layer normalization, a linear layer with the Tanh activation, and a dropout layer. Finally, we update the six decomposed features and fuse them to train the ConFEDE model with the following multi-task learning objective function:

$$\mathcal{L}_{\texttt{all}} = \mathcal{L}_{\texttt{pred}} + \beta_{\texttt{uni}}\mathcal{L}_{\texttt{uni}} + \beta_{\texttt{cl}}\mathcal{L}_{\texttt{cl}},$$

where $\mathcal{L}_{\texttt{pred}}$ is the multimodal prediction loss, $\mathcal{L}_{\texttt{uni}}$ represents the unimodal prediction loss and $\mathcal{L}_{\texttt{cl}}$ represents the contrastive loss. $\beta_{\texttt{cl}}$ and $\beta_{\texttt{uni}}$ are hyper-parameters that balance the contribution of each regularization component to the overall loss $\mathcal{L}_{\texttt{all}}$. We describe each loss term as follows.

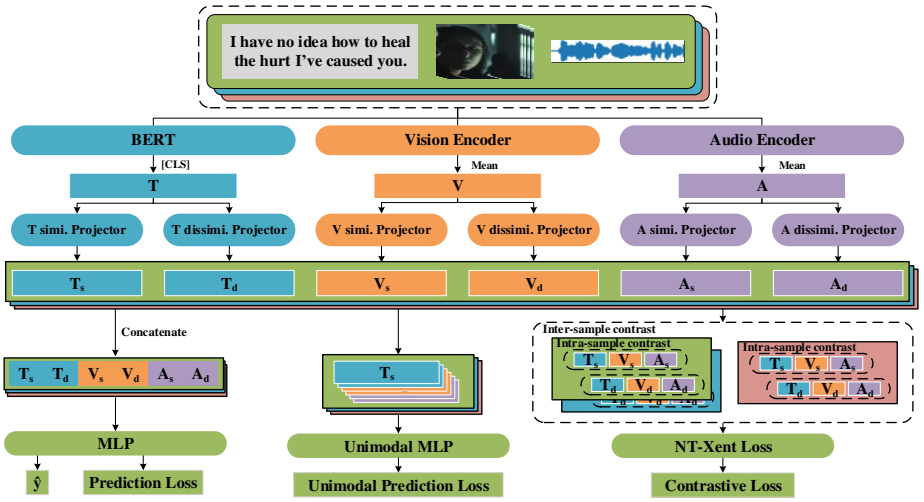

Figure 1: The structure of the ConFEDE framework. $\mathbf{T}_s$ and $\mathbf{T}_d$ represent the projected similarity and dissimilarity text features. Similarly, $\mathbf{V}_s$, $\mathbf{V}_d$, $\mathbf{A}_s$ and $\mathbf{A}_d$ are the projected similarity and dissimilarity features of vision and audio modalities, respectively.

$\mathcal{L}_{\texttt{pred}}$-Multimodal Prediction Loss. We use a multilayer perceptron (MLP) with the ReLU activation function as the classifier to get the final predictive result (i.e., $\hat{y}$ in Figure 1). We concatenate all 6 decomposed modality features to obtain the input to the classifier, $[\mathbf{T}_s^i; \mathbf{T}_d^i; \mathbf{V}_s^i; \mathbf{V}_d^i; \mathbf{A}_s^i; \mathbf{A}_d^i]$, where $[\cdot;\cdot]$ denotes the concatenation of two vectors. Denote the set of samples in a batch as B. For a given sample $i \in B$, let its prediction from the classifier be $\hat{y}_m^i$, we calculate the multimodal prediction loss by mean squared error:

$$\hat{y}_m^i = \texttt{MLP}([\mathbf{T}_s^i; \mathbf{V}_s^i; \mathbf{A}_s^i; \mathbf{T}_d^i; \mathbf{V}_d^i; \mathbf{A}_d^i]),$$

$$\mathcal{L}_{\texttt{pred}} = \frac{1}{n}\sum_{i=1}^{n}(y_m^i - \hat{y}_m^i)^2,$$

where n is the number of samples in a batch and y_m^i is the multimodal label.

$\mathcal{L}_{\texttt{uni}}$-Unimodal Prediction Loss. For each sample i, we also feed the 6 decomposed features $[\mathbf{T}_s^i, \mathbf{V}_s^i, \mathbf{A}_s^i, \mathbf{T}_d^i, \mathbf{V}_d^i, \mathbf{A}_d^i]$ into a weight-shared MLP classifier separately to get the 6 predictions denoted by the vector $\hat{\boldsymbol{u}}^i$. Specifically, we compute the unimodal prediction loss by:

$$\hat{\boldsymbol{u}}^i = \texttt{MLP}([\mathbf{T}_s^i, \mathbf{V}_s^i, \mathbf{A}_s^i, \mathbf{T}_d^i, \mathbf{V}_d^i, \mathbf{A}_d^i]),$$

$$\boldsymbol{u}^i = [y_m^i, y_m^i, y_m^i, y_t^i, y_v^i, y_a^i],$$

$$\mathcal{L}_{\texttt{uni}} = \frac{1}{n}\|\boldsymbol{u}^i - \hat{\boldsymbol{u}}^i\|_2^2,$$

where the vector $\boldsymbol{u}^i = [y_m^i, y_m^i, y_m^i, y_t^i, y_v^i, y_a^i]$ represents the ground-truth labels for unimodal prediction. In other words, each decomposed feature is regularized to perform prediction individually.

Note that the similarity features $\mathbf{T}_s^i, \mathbf{V}_s^i, \mathbf{A}_s^i$ are mapped through the MLP to predict the multimodal label y_m^i, whereas the dissimilarity features $\mathbf{T}_d^i, \mathbf{V}_d^i, \mathbf{A}_d^i$ are mapped through the MLP to predict modality-specific labels y_t^i, y_v^i, y_a^i (if available). When modality-specific labels are not available, the dissimilarity features $\mathbf{T}_d^i, \mathbf{V}_d^i, \mathbf{A}_d^i$ will also be used to predict multimodal label y_m^i. The rationale behind this design is that we let the similarity features capture the consistent information shared across different modalities via the overall multimodal label for the sample, while the dissimilarity features can retain modality-specific information represented by unimodal labels.

$\mathcal{L}_{\texttt{cl}}$-Contrastive Loss. We further regularize the learning through Contrastive Feature Decomposition in one simple joint contrastive loss that contrasts (1) similar samples against dissimilar samples; (2) similarity features against dissimilarity features within a sample. The contrastive loss is denoted as:

$$\mathcal{L}_{\texttt{cl}} = \frac{1}{n}\sum_{i=1}^{n}\ell_{\texttt{cl}}^i,$$

where $\ell_{\texttt{cl}}^i$ is the contrastive loss of sample i, the detailed derivation of which will be given in the following subsection.

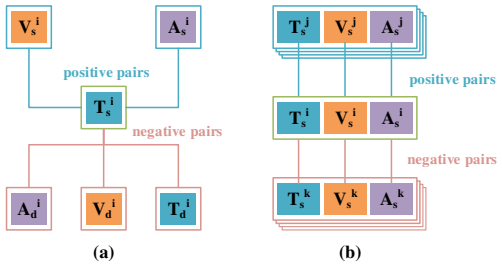

Figure 2: The pairing policy of decomposed modality features. (a) represents the intra-sample pairing; (b) represents the pairing between samples.

3.2 Contrastive Feature Decomposition

We unify intra-sample and inter-sample contrastive learning into one simple NT-Xent contrastive loss framework (Chen et al., 2020) to conduct both modality representation learning and modality decomposition simultaneously. The loss for sample i is given by

$$\ell_{\mathrm{cl}}^{i} = \sum_{(\boldsymbol{a},\boldsymbol{p})\in\mathcal{P}^i} -\log \frac{\exp(\mathtt{sim}(\boldsymbol{a},\boldsymbol{p})/\tau)}{\sum_{(\boldsymbol{a},\boldsymbol{k})\in\mathcal{N}^i\cup\mathcal{P}^i}\exp(\mathtt{sim}(\boldsymbol{a},\boldsymbol{k})/\tau)},$$

where $(\boldsymbol{a},\boldsymbol{p})$ and $(\boldsymbol{a},\boldsymbol{k})$ denote a pair of decomposed feature vectors either within a sample, e.g., $(\mathbf{T}_s^i, \mathbf{V}_s^i)$, $(\mathbf{T}_s^i, \mathbf{A}_d^i)$, or across different samples, e.g., $(\mathbf{T}_s^i, \mathbf{T}_s^j)$. The sets $\mathcal{P}^i$ and $\mathcal{N}^i$ are given by

$$\mathcal{P}^i = \mathcal{P}^i_{\mathtt{intra}} \cup \mathcal{P}^i_{\mathtt{inter}},$$

$$\mathcal{N}^i = \mathcal{N}^i_{\mathtt{intra}} \cup \mathcal{N}^i_{\mathtt{inter}}.$$

Here $\mathcal{P}^i$ is the positive pair set that includes both intra-sample positive pairs $\mathcal{P}^i_{\mathtt{intra}}$ and inter-sample positive pairs $\mathcal{P}^i_{\mathtt{inter}}$, while $\mathcal{N}^i$ is the negative pair set that consists of both intra-sample negative pairs $\mathcal{N}^i_{\mathtt{intra}}$ and inter-sample negative pairs $\mathcal{N}^i_{\mathtt{inter}}$. Note that $(\boldsymbol{a},\boldsymbol{p})$ is a positive pair in $\mathcal{P}^i$, and $(\boldsymbol{a},\boldsymbol{k})$ is a pair in $\mathcal{P}^i$ or $\mathcal{N}^i$.

Specifically, we use the six decomposed features ($\mathbf{T}_s$, $\mathbf{V}_s$, $\mathbf{A}_s$, $\mathbf{T}_d$, $\mathbf{V}_d$, $\mathbf{A}_d$) to form intra-sample positive/negative pairs, as shown in Figure 2 (a), with $\mathcal{P}^i_{\mathtt{intra}}$ and $\mathcal{N}^i_{\mathtt{intra}}$ given by

$$\begin{aligned}\mathcal{P}^i_{\mathtt{intra}} =& \{(\mathbf{T}_s^i, \mathbf{V}_s^i), (\mathbf{T}_s^i, \mathbf{A}_s^i)\} \\ &\cup \{(\mathbf{T}_s^j, \mathbf{V}_s^j), (\mathbf{T}_s^j, \mathbf{A}_s^j) \\ &\quad |j \in \mathtt{Neighbor}^i \cup \mathtt{Outlier}^i\},\end{aligned}$$

$$\begin{aligned}\mathcal{N}^i_{\mathtt{intra}} =& \{(\mathbf{T}_s^i, \mathbf{T}_d^i), (\mathbf{T}_s^i, \mathbf{V}_d^i), (\mathbf{T}_s^i, \mathbf{A}_d^i)\} \\ &\cup \{(\mathbf{T}_s^j, \mathbf{T}_d^j), (\mathbf{T}_s^j, \mathbf{V}_d^j), (\mathbf{T}_s^j, \mathbf{A}_d^j) \\ &\quad |j \in \mathtt{Neighbor}^i \cup \mathtt{Outlier}^i\},\end{aligned}$$

where $\mathtt{Neighbor}^i$ and $\mathtt{Outlier}^i$ represent the similar samples and dissimilar samples for the sample i, respectively, to enlarge the scope of the contrast, the detail of which is given in Algorithm 1 that will be explained subsequently.

Note that instead of treating all modalities equally as in other contrastive learning schemes, here we choose the text similarity feature $\mathbf{T}_s^i$ as an anchor, such that the visual and audio similarity features $\mathbf{V}_s^i$ and $\mathbf{A}_s^i$ are pushed closer to $\mathbf{T}_s^i$, while in the meantime, the dissimilarity features in all modalities are pushed away from $\mathbf{T}_s^i$. This is due to the observation that multimodal sentiment analysis is still largely centered around text information. Although other modalities can provide additional information to assist with sentiment prediction, they may also introduce more noise than text. Therefore, unlike other work, we avoid using visual/audio similarity features as anchors, which may bring noise into contrastive learning and confuse model training.

We now describe the data sampler shown in Algorithm 1 that retrieves similar samples for a given sample based on both multimodal features and multimodal labels to perform supervised contrastive learning across samples. Specifically, the sampling procedure can be divided into two steps.

First, given the dataset D that contains $|D|$ samples, for each sample pair (i, j) in D, we calculate the cosine similarity score between them:

$$\mathcal{C}^{i,j} = \mathtt{sim}([\mathbf{T}^i; \mathbf{V}^i; \mathbf{A}^i], [\mathbf{T}^j; \mathbf{V}^j; \mathbf{A}^j]),$$

where $\mathtt{sim}(\boldsymbol{w},\boldsymbol{v}) = \boldsymbol{w}^T\boldsymbol{v}/||\boldsymbol{w}|| \cdot ||\boldsymbol{v}||$ denotes the cosine similarity between two vectors $\boldsymbol{w}$ and $\boldsymbol{v}$. And $\mathbf{T}$, $\mathbf{V}$, and $\mathbf{A}$ (in Figure 1) are the output of BERT, vision and audio encoders, respectively.

Second, we retrieve candidate similar/dissimilar sample sets for each sample. For each sample i, we sort samples that have the same multimodal label y_m^i according to the similarity scores in ascending order as a candidate similar sample set S_0^i. In contrast, we sort samples that have labels other than y_m^i as a candidate dissimilar sample set S_1^i.

Two similar samples with high cosine similarity scores from S_0^i are randomly selected to form inter-sample positive pairs with sample i, which is denoted as $\mathtt{Neighbor}^i$. Four dissimilar samples from S_1^i are selected to form inter-sample negative pairs. We denote them as $\mathtt{Outlier}^i$ in which two samples $\mathtt{Outlier}_1^i$ have low cosine similarity scores and the other two samples $\mathtt{Outlier}_2^i$ have high cosine similarity scores.

Usually, we tend to select the samples in $\texttt{Neighbor}^i$ and $\texttt{Outlier}_1^i$ to form positive and negative pairs with sample i, respectively. However, samples in $\texttt{Outlier}_2^i$ have different labels but similar semantic information to sample i, making them hard to distinguish from sample i. Therefore, we additionally add these samples to $\texttt{Outlier}^i$ to specifically handle this issue by contrastive learning.

Based on the samples retrieved by Algorithm 1 and the pairing strategy shown in Figure 2 (b), the inter-sample positive/negative pairs for sample i are given by:

$$\mathcal{P}^i_{\texttt{inter}} = \{(\mathbf{T}^i_s, \mathbf{T}^j_s), (\mathbf{V}^i_s, \mathbf{V}^j_s), (\mathbf{A}^i_s, \mathbf{A}^j_s) \mid j \in \texttt{Neighbor}^i\},$$

$$\mathcal{N}^i_{\texttt{inter}} = \{(\mathbf{T}^i_s, \mathbf{T}^k_s), (\mathbf{V}^i_s, \mathbf{V}^k_s), (\mathbf{A}^i_s, \mathbf{A}^k_s) \mid k \in \texttt{Outlier}^i\}.$$

Notably, our data sampler enables contrastive learning across samples through decomposed modality features without data augmentation. This contrasts original contrastive learning in image classification, which obtains positive pairs by augmentation applied to images. Moreover, we only use similarity features to obtain inter-sample pairing since the similarity features of similar samples in the same class should be close while the similarity features of samples in different classes should be far apart.

Algorithm 1: Data Sampling Algorithm

Input: Dataset D with the corresponding features T, V, A and multimodal labels y_m.
Output: $\texttt{Neighbor}^i$, $\texttt{Outlier}^i$ for every $i \in D$
Define: $\texttt{sim}(\boldsymbol{w}, \boldsymbol{v}) = \boldsymbol{w}^T\boldsymbol{v}/||\boldsymbol{w}|| \cdot ||\boldsymbol{v}||$
for every $(i, j) \in D$ **do**
  Compute the cosine similarity score:
  $$\mathcal{C}^{i,j} = \texttt{sim}([\mathbf{T}^i; \mathbf{V}^i; \mathbf{A}^i], [\mathbf{T}^j; \mathbf{V}^j; \mathbf{A}^j]),$$
end
Define:
 $\texttt{argsort}(X) = \texttt{indices sort X ascendingly}$
Let $|D| = \texttt{length of } D$, $z = \frac{|D|}{4}$.
for every sample $i \in D$ **do**
  Retrieve the similar sample set S_0^i:
  $$S_0^i = \texttt{argsort}(\{\mathcal{C}^{i,j} | j : y_m^j = y_m^i\});$$
  Retrieve the dissimilar sample set S_1^i:
  $$S_1^i = \texttt{argsort}(\{\mathcal{C}^{i,j} | j : y_m^j \neq y_m^i\}),$$
  Randomly select two samples from the last z elements of S_0^i as $\texttt{Neighbor}^i$;
  Randomly select two samples from the first z elements of S_1^i as $\texttt{Outlier}_1^i$;
  Randomly select two samples from the last z elements of S_1^i as $\texttt{Outlier}_2^i$;
  $$\texttt{Outlier}^i = \texttt{Outlier}_1^i \cup \texttt{Outlier}_2^i.$$
end

4 Experiments

We mainly evaluate ConFEDE on CH-SIMS (Yu et al., 2020), since it has unimodal labels, which can best meet the design of ConFEDE. To justify the effectiveness of ConFEDE when unimodal labels are unavailable, we further test ConFEDE on the MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018b), which are two English MSA datasets. However, they can not best test the performance of ConFEDE.

We compare our methods with the state-of-the-art baselines in Table 1 and 2: LF-DNN (Yu et al., 2020), MFN (Zadeh et al., 2018a), LMF (Liu et al., 2018), TFN (Zadeh et al., 2017), MulT (Tsai et al., 2019), MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), HyCon (Mai et al., 2021) and Self-MM (Yu et al., 2021). For a fair comparison, the methods which only report the results of a single run and have no valid official code released for reproduction are not selected. A detailed introduction can be found in the supplementary material. The detailed experimental settings are introduced in Appendix C.

4.1 Evaluation Metrics

Following the previous works (Yu et al., 2020, 2021; Rahman et al., 2020; Hazarika et al., 2020), we report our results in (multi-class) classification and regression with the average of 5 runs of different seeds. For classification, we report the multi-class accuracy and weighted F1 score. We calculate the accuracy of 2-class prediction (Acc-2), 3-class prediction (Acc-3), and 5-class (Acc-5) prediction for CH-SIMS and the accuracy of 2-class prediction and 7-class prediction (Acc-7) for MOSI and MOSEI. Besides, Acc-2 and F1-score of MOSI and MOSEI have two forms: negative/non-negative (non-exclude zero) (Zadeh et al., 2017; Yu et al., 2021) and negative/positive (exclude zero) (Tsai et al., 2019; Yu et al., 2021). For regression, we report Mean Absolute Error (MAE) and Pearson correlation (Corr). Except for MAE, higher values

Model	CH-SIMS					
	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
LF-DNN	78.87	79.87	66.91	41.62	0.420	0.612
MFN(A)	77.9	77.88	65.73	39.47	0.435	0.582
LMF	77.77	77.88	64.68	40.53	0.441	0.576
TFN	78.38	78.62	65.12	39.30	0.432	0.591
MulT(A)	78.56	79.66	64.77	37.94	0.453	0.561
Self-MM	80.04	80.44	65.47	41.53	0.425	0.595
Plain	80.79	80.56	68.49	42.98	0.393	0.664
Inter	81.14	81.29	68.84	45.64	**0.381**	0.656
Intra	81.88	81.84	69.36	43.24	0.391	**0.668**
ConFEDE	**82.23**	**82.08**	**70.15**	**46.30**	0.392	0.637

Table 1: Results on CH-SIMS. For each model, (A) means the model utilized the aligned data

indicate better performance for all metrics.

4.2 Results

The performance comparison of all methods on CH-SIMS, MOSI, and MOSEI is summarized in Table 1 and Table 2. The scores of the proposed method and its variations are the averages of 5 runs. The performances of all other baselines, except for MAG-BERT, have been sourced from published papers or official repositories[1].

On the CH-SIMS dataset, our proposed method outperforms all baselines on all metrics. We achieve superior performance compared to the best baseline model, Self-MM, with an improvement of 2.19% on acc-2 and 1.64% on F1 scores. Additionally, the proposed model demonstrates exceptional ability in multi-class classification, outperforming the best baseline by 4.68% on acc-3 and 4.77% on acc-5.

As seen in the results, our proposed method, ConFEDE, consistently outperforms all other baselines on the CH-SIMS dataset. The superior classification performance demonstrates that our designed learning method is more effective than the compared methods. Our method, ConFEDE, effectively distinguishes similarity and dissimilarity information between modalities, providing clearer modality features to the downstream classifier for improved prediction. Additionally, the significant improvement in MAE and Corr further highlights the ability of our model to better understand the CH-SIMS dataset than the other baselines.

To further evaluate the effectiveness of our proposed method, ConFEDE, we trained our models on the MOSI and MOSEI datasets without unimodal labels. Instead, we used their multimodal labels for compatibility. The results are presented in Table 2. On the MOSI dataset, our method outperforms all other baselines in both the negative/non-negative (NN) setting and negative/positive (NP) setting for acc-2 and F1 metrics. Additionally, our acc-7 and MAE metrics surpass most of the baselines. For the MOSEI dataset, our ConFEDE method outperforms all baselines in all metrics except for the NN Acc-2 and F1 score. Furthermore, our MAE is significantly lower than all baselines, reaching 0.522.

It is worth noting that our models perform much better in NP acc-2 than NN acc-2 for MOSEI, as shown in Table 2. This is because the NN acc-2 setting is generally more challenging than the NP acc-2 setting, as it places more pressure on a model to classify data samples with a regression label of 0. Specifically, if there are two samples with a regression label of 0, when predicted by a regression model, the results might be -0.01 and 0.01. As the value range of "Neutral" is [-0.5,0.5) in MOSI and (-0.1,0.1] in SIMS, these two samples should be classified as "Neutral". However, in NN settings, they will be classified into two different classes, resulting in a worse acc-2. In contrast, with the NP setting, all "Neutral" samples are abandoned, resulting in a better acc-2.

In contrast, our method shows better performance in both NN and NP settings on MOSI when compared to other models. The Acc-7, MAE and Corr are also better or comparable to most baselines.

4.3 Ablation Study and Analysis

To evaluate the impact of our proposed structures, we conducted an ablation study on our proposed method by removing inter-sample contrastive learning and intra-sample contrastive learning. The results are shown in Table 1. "Plain" represents the model without contrastive learning method, "Inter" represents the model with inter-sample contrastive learning only, and "Intra" represents the model with intra-sample contrastive learning and unimodal prediction as a sub-task.

The experiment shows that all three models perform worse than the original model. Among the three models, the plain setting has the lowest performance. Both intra-sample contrastive learning and inter-sample contrastive learning provide positive impacts on performance. Compared with Plain,

[1]https://github.com/thuiar/MMSA/blob/master/results/result-stat.md

Model	MOSI					MOSEI				
	Acc-2	F1	Acc-7	MAE	Corr	Acc-2	F1	Acc-7	MAE	Corr
LF-DNN	77.52/78.63	77.46/78.63	34.52	0.955	0.658	80.60/82.74	80.85/82.52	50.83	0.58	0.709
MFN(A)	77.4/-	77.3/-	34.1	0.965	0.632	78.94/82.86	79.55/82.85	51.34	0.573	0.718
LMF	-/82.5	-/82.4	33.2	0.917	0.695	80.54/83.48	80.94/83.36	51.59	0.576	0.717
TFN	-/80.8	-/80.7	34.9	0.901	0.698	78.50/81.89	78.96/81.74	51.60	0.573	0.714
MulT(A)	-/83.0	-/82.8	40.0	0.871	0.698	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA(A)	81.8/83.4	81.7/83.6	42.3	0.783	0.776	83.6/85.5	83.8/85.3	52.2	0.555	0.756
MAG-BERT	82.13/83.54	81.12/83.58	41.43	0.790	0.766	79.86/86.86	80.47/83.88	50.41	0.583	0.741
HyCon	-/85.2	-/85.1	46.6	0.713	0.790	-/85.4	-/85.6	52.8	0.601	0.776
Self-MM	83.44/85.46	83.36/85.43	**46.67**	**0.708**	**0.796**	**83.76**/85.15	**83.82**/84.90	53.87	0.531	0.765
ConFEDE	**84.17/85.52**	**84.13/85.52**	42.27	0.742	0.784	81.65/**85.82**	82.17/**85.83**	**54.86**	**0.522**	**0.780**

Table 2: Results on MOSI and MOSEI. The settings and results of all baselines are same with Table 1. In Acc-2 and F1 score, the left of the "/" corresponds to "negative/non-negative" and the right corresponds to "negative/positive".

Model	Acc-2	F1	Acc-3	Acc-5	MAE	Corr
Intra	**81.88**	**81.84**	**69.36**	**43.24**	0.391	0.668
M-label	80.96	81.08	68.84	41.92	0.389	**0.671**
-uni	81.71	81.71	68.97	42.27	0.389	0.669
-cl	80.92	81.06	68.84	42.71	**0.383**	0.668
+full	81.58	81.42	**69.36**	42.45	0.391	0.669

Table 3: The Ablation study results of Intra on CH-SIMS. Each score is the average of 5 runs.

by using text feature as the anchor, Intra filters out noise (useless information for sentiment analysis) in the vision and audio modality, leading to better prediction. This is also the reason why it reaches better acc-2 accuracy than both the other models. Since the acc-2 metric in CH-SIMS follows the negative/non-negative setting, a feature with lower noise helps the classifier make a more precise prediction value, making it easier to classify the 0-labeled samples. This also explains why we achieve better NN Acc-2 performance than all baselines on MOSI.

For the inter-sample contrastive learning method, by learning the common and different information between samples, "Inter" performs better on multiclass classification. The result of Inter on CH-SIMS shows great improvements on both acc-5 and MAE compared with the other two models, which proves that Acc-5 and regression performance benefits more from "Inter". This can also explain why we have a lower NN Acc-2 performance on MOSEI. Since MOSEI is much larger than MOSI and CH-SIMS, it introduces more noise in each modality, the contrastive feature decomposition learning needs more epochs and a smaller learning rate to separate the useful information from noise. Meanwhile, inter-sample contrastive learning is more efficient on MOSEI. With the larger amount of samples, it is much easier for the sampler to find the most similar and dissimilar samples with the given sample, from which the model can understand the difference between samples better. Thus, ConFEDE can reach higher Acc-7 and regression performance than all other baselines on MOSEI.

To further evaluate the effectiveness of the contrastive feature decomposition method, we conducted an ablation study on Intra using the CH-SIMS dataset. As presented in Table 3, we created three variations of Intra: 1) Intra with only multimodal labels for unimodal prediction (M-label); 2) Intra without the unimodal prediction component (-uni); 3) Intra without the similarity-dissimilarity learning method (-cl); and 4) Intra that uses all similarity features as anchors (+full), which utilizes T_s, V_s, and A_s as anchors instead of T_s only.

The results in the table demonstrate that all variations resulted in a decrease in performance compared to the original Intra in classification matrices. Both the intra-sample contrastive learning and the unimodal prediction task can regularize the learned representation, resulting in clearer information that aids the classifier in understanding the sample better. However, the "+full" setting introduces more noise by also using V_s and A_s as anchors, which confuses the model and diminishes the denoising ability of the contrastive feature decomposition learning.

5 Conclusion

In this paper, we propose a novel method for multimodal sentiment analysis (MSA) called ConFEDE. The ConFEDE framework is based on contrastive feature decomposition, which utilizes a unified contrastive training loss to capture the consistency and difference across modalities and samples. This approach allows for the simultaneous learning of modality decomposition within each sample and su-

pervised contrastive learning across samples. Our proposed method is mainly evaluated on CH-SIMS. The result shows that the proposed method significantly outperforms many state-of-the-art multimodal sentiment analysis methods. We further conduct an extensive experiment on MOSI and MOSEI to test the capability of ConFEDE when no unimodal label is available, where our method achieves better performance than state-of-the-art methods on a number of performance metrics.

Limitations

While our proposed ConFEDE method has shown promising results in multimodal sentiment analysis, there are some limitations to consider. Firstly, our method is designed for multimodal sentiment analysis that includes three modalities: vision, audio, and text. The performance of the model when one of these modalities is missing is not considered. Additionally, as the number of training samples increases, our custom-designed sampling method may require more processing time. However, the similarity calculation can be pre-processed between the unimodal training stage and the multimodal training stage (as outlined in Appendix C). Therefore, it may not consume a significant amount of time.

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618.

Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126. PMLR.

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.

Rongrong Ji, Fuhai Chen, Liujuan Cao, and Yue Gao. 2019. Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning. *IEEE Transactions on Multimedia*, 21(4):1062–1075.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065.

Tianliang Liu, Junwei Wan, Xiubin Dai, Feng Liu, Quanzeng You, and Jiebo Luo. 2019. Sentiment recognition for short annotated gifs using visual-textual fusion. *IEEE Transactions on Multimedia*, 22(4):1098–1110.

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.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2021. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *arXiv preprint arXiv:2109.01797*.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for*

Computational Linguistics. Meeting, volume 2020, page 2359. NIH Public Access.

Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999.

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 conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. 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.

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.

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, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. 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.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. 2022. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*, pages 1–1.

A Datasets

Table 4 shows the statistics of these datasets.

CH-SIMS. The CH-SIMS dataset is a Chinese multimodal dataset that contains 2281 refined video segments. Each sample has one multimodal label and three unimodal labels with sentiment scores ranging from -1 (strongly negative) to +1 (strongly positive).

Dataset	#Train	#Valid	#Test	#Total	Language
CH-SIMS	1368	456	457	2281	Chinese
MOSI	1284	229	686	2199	English
MOSEI	16326	1871	4659	22856	English

Table 4: The statistics of CH-SIMS, MOSI and MOSEI.

MOSI. The MOSI dataset is a popular dataset with three modalities (*i.e.*, text, video and audio) specially designed for sentiment analysis. It was collected from 93 YouTube videos where a speaker expresses opinions about movies. MOSI contains 2199 utterance-video clips. Each clip was annotated with a sentiment score ranging from -3 (strongly negative) to +3 (strongly positive).

MOSEI. The MOSEI dataset is the larger version of MOSI that contains 22856 annotated video segments over 250 different topics. Same as MOSI, each clip has a sentiment score between -3 (strongly negative) to +3 (strongly positive).

B Baselines

LF-DNN. The Later Fusion DNN (LF-DNN) simply concatenates unimodal features that are extracted by unimodal subnets for sentiment inference(Yu et al., 2020)

MFN. The Memory Fusion Network (MFN) (Zadeh et al., 2018a) first learns view-specific interactions by LSTM, then learns the cross-view interactions by an attention network, and finally summarizes through time with a Multi-view Gated Memory. The outputs of MFN are concatenated as the final representations.

LMF. The Low-rank Multimodal Fusion (LMF) method (Liu et al., 2018) leverages low-rank tensor to perform multimodal fusion efficiently.

TFN. The Tensor Fusion Network (TFN) (Zadeh et al., 2017) consists of 1) modality embedding subnetworks that take unimodal features as input and

output rich encodings after a neural network, 2) tensor fusion layer that models the unimodal, bimodal and trimodal interactions using outer-product, 3) sentiment inference subnetwork that performs sentiment inference.

MulT. The Multimodal Transformer (MulT) (Tsai et al., 2019) utilizes the directional pairwise cross-modal attention to learn the interactions between multimodal sequences, and latently adapt streams from one modality to another.

MISA. MISA (Hazarika et al., 2020) is a multimodal framework that learns modality-invariant and modality-specific representations for each modality. The learning process is optimized by a combination of losses including similarity loss, orthogonal loss, reconstruction loss, and task prediction loss.

MAG-BERT. The Multimodal Adaptation Gate for Bert (MAG-BERT) (Rahman et al., 2020) is developed by applying multimodal adaptation gate at different layers of the BERT backbone. We reproduce this method and take the average result of 5 runs for a fair comparison with our method[2].

HyCon. Hybrid Contrastive Learning of Tri-modal Representation (HyCon) (Mai et al., 2021) is developed based on contrastive learning method. It focus on the interaction between modalities and classes.

Self-MM. Self-MM (Yu et al., 2021) first utilizes a self-supervised label generation module to obtain unimodal labels, then joint learn the multimodal and unimodal representations based on the multimodal label and generated unimodal labels.

C Experiments

C.1 Experimental Settings

Here we briefly introduce the detailed settings of our experiments. All experiments were conducted on a single NVIDIA RTX 3090 GPU. ConFEDE has less than 200 million parameters for all implementations. The training is consist of a unimodal training stage and a multimodal training stage.

In the unimodal training stage, we employ "bert-base-chinese"[3] for CH-SIMS and "bert-base-uncased"[4] for MOSI and MOSEI. we fine-tune both Chinese BERT and English BERT with a learning rate of 0.00001 with a batch size of 64, and train 150 epochs to ensure convergence. We then employ transformer encoders as our Vision Encoder and Audio Encoder in Figure 1. Specifically, for CH-SIMS and MOSI, we use two single-layer transformer encoders (Vaswani et al., 2017) on each of them to extract the audio and the visual information respectively. For MOSEI, we use 3 transformer layers to build each decoder, since MOSEI is much larger than the other two. All vision encoders and audio encoders are trained for 300 epochs with the learning rate equals 0.0001 and batch size equals 128.

In the multimodal stage, we train ConFEDE for MSA with the encoders obtained above. The loss ratio is set to be $\beta_{cl} = 0.1$ and $\beta_{uni} = 0.01$. For CH-SIMS and MOSI, we train ConFEDE for 50 epochs with the learning rate of 0.0001. We set the batch size to 32 for CH-SIMS and 16 for MOSI. For MOSEI, the model is trained with a batch size of 4 for 25 epochs. The learning rate is set to 0.00005.

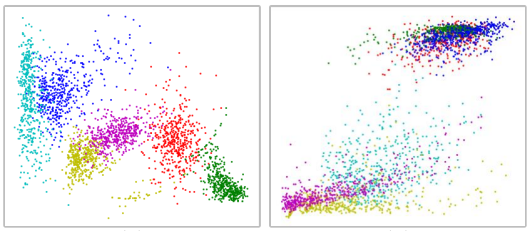

Figure 3: The 2D projections of all six similarity-dissimilarity features extracted and projected by ConFEDE model on the test set of CH-SIMS: (a) represents the decomposed features without ConFEDE; (b) represents the decomposed features with ConFEDE. The colors red, green, blue, cyan, yellow and magenta represent T_s, V_s, A_s, T_d, V_d and A_d, respectively.

C.2 Visualization

Figure 3 shows the 2D projections of the six decomposed features of all test samples on CH-SIMS, where (a) is the six decomposed features without ConFEDE and (b) shows these features with ConFEDE. From it, we can observe that the similarity features (i.e., T_s in red, V_s in green and A_s in blue) become closer to each other while the dissimilarity features (i.e., T_d in cyan, V_d in yellow and A_d in magenta) become further away from their corresponding similarity features, indicating the effectiveness of ConFEDE to learn the consistency and incongruity between modalities.

[2]We select the model with the best validation performance to evaluate the test set.

[3]https://huggingface.co/bert-base-chinese

[4]https://huggingface.co/bert-base-uncased

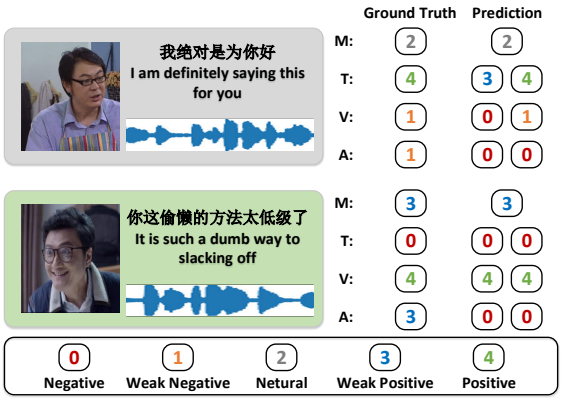

Figure 4: Two real-case examples from the test set of CH-SIMS. "M" represents the multimodal sentiment label of the sample, while "T", "V" and "A" represent the unimodal sentiment labels. For unimodal labels in "Prediction", the first column represents the unimodal prediction using similar features, while the second column represents the unimodal prediction using dissimilar features.

C.3 Real Case

Figure 4 provides two real cases from CH-SIMS. From the figure, we can observe that the multimodal sentiment can be much different from the unimodal sentiment labels. In the first example, the text is positive by having the word "good". However, the man is having a serious face and his voice sounds angry, which indicates the man is actually explaining what he has said, making the overall sentiment neutral. In the second example, if only judging from the text modal, the sentiment is very negative by having words "dumb" and "slacking off". However, combining the happy face and the fun voice, we can understand the man is bantering with someone, which makes the multimodal sentiment of the sample a weak positive.

By understanding the unimodal sentiments, ConFEDE can make the correct prediction on both samples. However, as discussed in the main paper, the visual modality and audio modality are much noisier than the text modality. This is also why ConFEDE makes more precise prediction on text than on audio and vision for the two samples.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☑ A4. Have you used AI writing assistants when working on this paper?
use Grammarly to check grammar

B ☑ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.