electronics

MDPI

Article

Self-HCL: Self-Supervised Multitask Learning with Hybrid Contrastive Learning Strategy for Multimodal Sentiment Analysis

Youjia Fu [1], Junsong Fu [1,*], Huixia Xue [1] and Zihao Xu [2]

[1] College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China; youjia_fu@cqut.edu.cn (Y.F.); xue_xue@stu.cqut.edu.cn (H.X.)
[2] Liangjiang Institute of Artificial Intelligence, Chongqing University of Technology, Chongqing 401135, China; xu139@stu.cqut.edu.cn
* Correspondence: juns_fu@stu.cqut.edu.cn

Abstract: Multimodal Sentiment Analysis (MSA) plays a critical role in many applications, including customer service, personal assistants, and video understanding. Currently, the majority of research on MSA is focused on the development of multimodal representations, largely owing to the scarcity of unimodal annotations in MSA benchmark datasets. However, the sole reliance on multimodal representations to train models results in suboptimal performance due to the insufficient learning of each unimodal representation. To this end, we propose Self-HCL, which initially optimizes the unimodal features extracted from a pretrained model through the Unimodal Feature Enhancement Module (UFEM), and then uses these optimized features to jointly train multimodal and unimodal tasks. Furthermore, we employ a Hybrid Contrastive Learning (HCL) strategy to facilitate the learned representation of multimodal data, enhance the representation ability of multimodal fusion through unsupervised contrastive learning, and improve the model's performance in the absence of unimodal annotations through supervised contrastive learning. Finally, based on the characteristics of unsupervised contrastive learning, we propose a new Unimodal Label Generation Module (ULGM) that can stably generate unimodal labels in a short training period. Extensive experiments on the benchmark datasets CMU-MOSI and CMU-MOSEI demonstrate that our model outperforms state-of-the-art methods.

Keywords: contrastive learning; feature optimization; multitask learning; multimodal sentiment analysis

Citation: Fu, Y.; Fu, J.; Xue, H.; Xu, Z. Self-HCL: Self-Supervised Multitask Learning with Hybrid Contrastive Learning Strategy for Multimodal Sentiment Analysis. *Electronics* **2024**, *13*, 2835. https://doi.org/10.3390/electronics13142835

Academic Editors: Leonidas Akritidis and Panayiotis Bozanis

Received: 24 June 2024
Revised: 15 July 2024
Accepted: 16 July 2024
Published: 18 July 2024

1. Introduction

The rapid development of neural network modeling has brought diverse techniques and methods to the field of human–computer interaction. Long Short-Term Memory Networks (LSTMs) [1] have effectively solved the limitations of traditional Recurrent Neural Networks (RNNs) [2] in dealing with long-term dependencies by introducing a gating mechanism, which is especially suitable for analyzing and predicting time series data. The Transformer model based on the self-attention mechanism is able to deal with long-range dependencies and is now widely used in various sequence modeling tasks. In addition, "Knowing knowledge: Epistemological study of knowledge in transformers [3]" investigates the role of neural models in human–computer interaction, thus providing new perspectives for understanding how neural networks facilitate knowledge exchange.

Multimodal sentiment analysis (MSA) plays a crucial role in the field of human–computer interaction and has become a hot research topic in recent years [4]. MSA has received much attention in recent years compared to traditional unimodal sentiment analysis methods, MSA has demonstrated significant advantages in terms of robustness, and it has made breakthroughs in processing social media data in particular. With the explosive growth of user-generated content, MSA has been used in a wide range of domains,

including social monitoring, consumer services, and the transcription of video content. By integrating information from different modalities, such as textual, audio, and visual data, this analytic approach is able to capture and parse the user's affective state more comprehensively, thus improving the accuracy and reliability of sentiment recognition.

Today, research in MSA mainly focuses on how to effectively learn joint representations. Researchers have evolved their work from tensor-based approaches [5] to approaches based on attention mechanisms [6,7], and they have continuously worked on designing modules that capture crossmodal information interactions and utilize multimodal representations to train models. However, relying solely on multimodal representations to train models often leads to suboptimal performance [8]. This is mainly due to the lack of unimodal annotations in the MSA benchmark dataset, thereby making it difficult for models to capture unimodal-specific information. As shown in Figure 1, uniform multimodal labels are not always appropriate for unimodal learning, which limits the model's ability to understand each unimodal state in depth. A number of attempts have been made by some researchers to solve this problem. Yu et al. [9] proposed the Self-MM, which calculates the distance between the modal representation and the category centroid to quantify the degree of similarity. Han et al. [10] designed the MMIM, which enhances the effect of multimodal fusion by increasing the mutual information between unimodal representations and the shared information between fusion embedding and unimodal representations. Furthermore, Hwang et al. [11] presented SUGRM using recalibration information to generate unimodal annotations with dynamically adjusted features. However, how to better learn unimodal feature representations and optimize multimodal feature representations in the absence of unimodal annotations remains to be further explored.

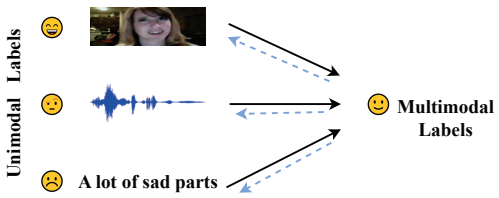


Figure 1. An example of unimodal labels and multimodal labels. The blue dotted lines represent the process of backpropagation.

In order to address the above problems, we designed an innovative Multimodal Sentiment Analysis framework called Self-HCL. The framework initially employs the Unimodal Feature Enhancement Module (UFEM) to optimize the learning of unimodal features. Specifically, the UFEM computes and assigns attentional weights to modal features in the channel and spatial dimensions by using the Convolutional Block Attention Module (CBAM) [12]. It then uses these weights to optimize the representation of unimodal features by finely tuning the original features and selectively reinforcing them through gating mechanisms and elemental multiplication. Next, the Sparse Phased Transformer (SPT) [13] is used to capture and integrate the final feature representations for each modality. In addition, Self-HCL integrates a Hybrid Contrastive Learning (HCL) strategy to optimize the representation learning process for multimodal data. On the one hand, we adopt the principle of Unsupervised Contrast Learning (UCL) [14], which enhances the extraction of interrelated information between the fused features and each unimodal modality through iterative operations so as to reveal the deep relationships between modalities and optimize the spatial layout of fused features. On the other hand, to address the problem of the scarcity of unimodal annotation data, we introduce a Supervised Comparative Learning (SCL) strategy. We map the features of different modalities into the same high-dimensional feature space to facilitate the aggregation of samples with the same emotion label in the embedding space while ensuring the differentiation of differently labeled samples. Finally, we improve the Unimodal Label Generation Module (ULGM) proposed by Hwang et al. [11]. We constructed a new UCL space based on it and combined with the properties of UCL, which enables the ULGM to output unimodal labels stably in a shorter period of time. The

improved ULGM not only fully utilizes the advantages of contrast learning in mining feature differences and uniqueness, but it also successfully overcomes the limitations encountered by Hwang et al. [11] in dealing with the modal feature similarity puzzle. To summarize, the primary contributions of this work are as follows:

- We construct a novel MSA framework called Self-HCL, which improves the identification of salient features in the absence of unimodal annotation using the UFEM and optimizes the features by combining the gating mechanism with element multiplication, which effectively improves the representation learning of unimodal features.
- A hybrid contrastive learning strategy is designed for the purpose of deep exploration of the fused multimodal features and the inherent relationship between each single modal feature and emotional labels.
- We propose an improved ULGM, which reveals the deep relationship between different modalities and optimizes the spatial distribution of modal features by constructing a new unsupervised contrastive learning space, thus achieving the stable generation of unimodal labels within a short training cycle.

2. Related Work

2.1. Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) is an approach for identifying and understanding emotions by analyzing speech, facial expressions, voice, music, and body movements. The discipline has advanced using publicly available datasets, including CMU-MOSI, CMU-MOSEI, and IEMOCAP [15]. There are three main MSA research directions: (1) Initially, multimodal fusion used techniques like tensor fusion networks [6] and low-rank multimodal fusion [16] with LSTM [1] to create high-dimensional tensors for integrating diverse data sources. (2) Modal interaction modeling [17] explores complex interactions between modalities using MCTN [18] and MulT [4], which enhance intermodal transformations using cyclic consistency loss and the Transformer architecture encoder/decoder, respectively. Sun et al. [19] offered deep normalized correlation analysis for improved intermodal consistency in high-dimensional nonlinear spaces. (3) Mode consistency and disparity techniques, which seek coherence and highlight discrepancies between modalities, have garnered attention. For example, Yu et al. [9] created a self-supervised learning module for label generation in multimodal and unimodal training tasks, thus minimizing mode differences. Han et al. [10] used mutual information in MSA and proposed a learning framework to preserve task-relevant information. In their model MISA [5], modal vectors were mapped into two spaces, and regularization was added to aid in learning shared and distinct modal properties.

2.2. Multitask Learning

Multitask learning is a key branch of machine learning that focuses on optimizing the connections between multiple related tasks simultaneously [20]. It falls under the migrating learning framework, which aims to extract and apply domain-specific knowledge from training data for various related tasks. In multitask learning, model parameters act as a sharing mechanism during training, thus allowing the model to extract common feature representations from different tasks to improve its generalization across various tasks. There are two main types of parameter sharing: soft sharing, where model parameters are adjusted for different tasks, and hard sharing, where fixed global parameters aid in learning all tasks. In the field of MSA, multitask learning has been widely used to integrate information from different modalities like text, speech, and image, thus leading to improved sentiment recognition and emotion analysis [21,22].

2.3. Contrastive Learning

Contrastive learning, based on the InfoNCE theory [23], uses a loss function to increase the mutual information between feature representations of the same object from different perspectives or conditions while reducing it between unrelated objects (negative

sample pairs). This approach helps the model develop more distinct feature representations. Recent methodologies like SimCLR [24] and MoCo [25] have advanced the practical applications and theoretical exploration of contrastive learning in computer vision, thus improving learning outcomes in unsupervised settings through data augmentation and queuing mechanisms. As deep learning techniques have evolved, contrastive learning has expanded beyond visual data like images to fields such as natural language processing and multimodal learning. It has been successful in extracting unified representations from various data types, including text, images, and audio. For example, Khosla et al. [26] extended contrastive learning by incorporating supervised information into the unsupervised framework, thus allowing for multiple positive samples to be associated with the same anchor sample. Moreover, Han et al. [10] enhanced contrastive learning by maximizing mutual information across different aspects of a single input instance, thus filtering and amplifying feature information relevant to the target task.

2.4. ULGM

Designed and developed by Yu et al. [9], ULGM aims to automatically generate unimodal labels for multimodal tasks. The module relies on the assumption that label differences between categories are directly related to differences in the distances of modal eigenvectors from category centers. Labels from unimodal data should align with those from multimodal fusion information. However, close interclass distances and indistinguishable category centers can lead ULGM to produce unstable or inaccurate labels, thus impacting learning stability and causing the model to converge to a local optimum. Hwang et al. [11] proposed an enhancement based on Yu et al. [9] to address this issue. The enhancement scheme generates unimodal labels based on distances between the feature space and label space. ULGM proposes that the distances between feature points in a semantic space are linked to the distances of their corresponding labels. By calculating feature distances using multimodal tag information, ULGM infers and generates unimodal tags. It considers offset size and direction, thereby determining the offset by comparing distances between multimodal and unimodal features with the maximum tag space distance and analyzing positive and negative tag center positions relative to multimodal features.

3. Approach

3.1. Problem Definition

MSA is a technique that combines multiple modal signals such as text, audio, and visual to accurately determine sentiment states. In this study, the input to the model is defined as I_s, where $s \in \{t, a, v\}$. And this composite input consists of three key components: textual modality, audio modality, and video modality. The core task of the model is to predict the corresponding sentiment intensity value $\hat{y}_m \in \mathbb{R}$ after receiving inputs such as I_s. To optimize the learning process of the model, in the training phase, we generate the corresponding labels $y_s \in \mathbb{R}$ for each modality separately. Although the model can produce multiple potential outputs, in practical applications, we only select $\hat{y}_m \in \mathbb{R}$ as the final sentiment prediction index.

3.2. Overall Architecture

Self-HCL facilitates the sharing of fundamental modal representations by incorporating multimodal tasks, unimodal activities, and hybrid contrastive learning tasks. When faced with problems that involve several modes of input and various types of unimodal tasks, we employ a hard sharing method to construct a shared underlying learning network. Figure 2 depicts the comprehensive structure of Self-HCL, thus showcasing how modal representation information may be efficiently exchanged and utilized across activities. In Figure 2, y_s is the unimodal annotation generated by ULGM based on the manually annotated multimodal labels y_m for supervised learning of the unimodal task. $\hat{y}_s$ and $\hat{y}_m$ are the predicted sentiments for the unimodal task and the multimodal task, respectively, where $s \in \{t, a, v\}$.

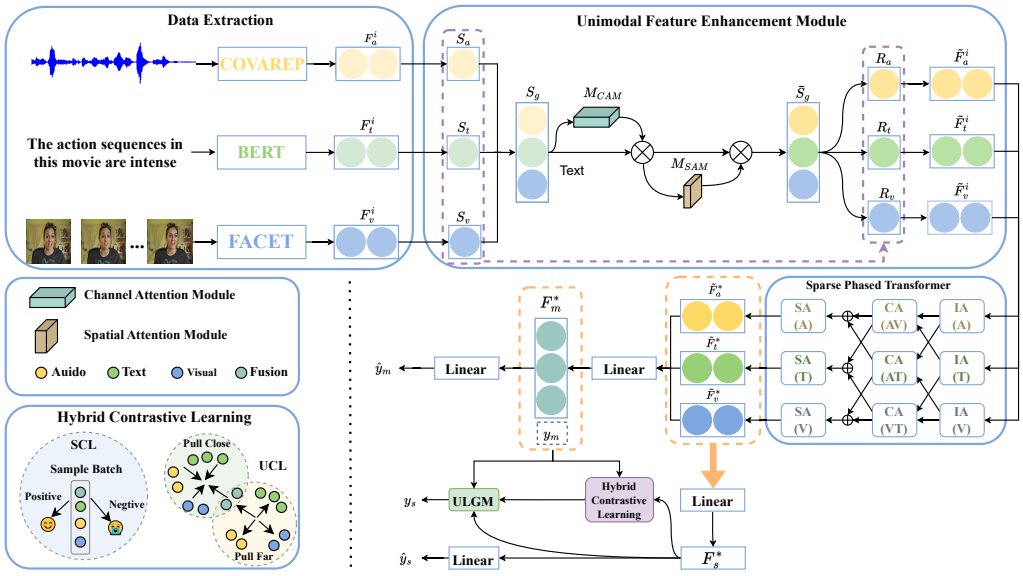


Figure 2. Overall architecture of Self-HCL.

3.3. Multimodal Task

For the multimodal task, we extract modality features F_s^i from pretrained BERT [27], COVAREP [28], and FACET [29] models for textual, acoustic, and visual input, respectively. Subsequently, the Unimodal Feature Enhancement Module (UFEM) is employed to optimize the extracted features for each modality type, and the Sparse Phased Transformer (SPT) is utilized to capture and integrate the final feature representation for each modality.

Unimodal Feature Enhancement Module: The UFEM primarily utilizes the Convolutional Block Attention Module (CBAM) [12], a specialized attention mechanism module designed for Convolutional Neural Networks (CNNs) [30], thus aiming to enhance the network's expressiveness and performance in processing visual tasks by strengthening the attention to key features. The CBAM comprises two primary modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM). Here, we show how the CBAM can be applied to the UFEM. The UFEM receives $F_s^i \in \mathbb{R}^{l_s \times d_s}$ as input, where l_s is the length of the sequence, and d_s is the modal feature dimension, and we squeeze the input along the sequence length using global average pooling:

$$S_s(d) = \frac{1}{l_s} \sum_{l=1}^{l_s} F_s^i(l, d) \tag{1}$$

where $s \in \{t, a, v\}$, and $d = 1, 2, .., ds$. The compression feature S_s is then connected and fed into a series of fully connected networks and ReLU to learn the global multimodal embedding S_g:

$$S_g = ReLU(W_g[S_t; S_a; S_v] + b_g) \tag{2}$$

where $[;]$ denotes the feature concatenate, W_z is a 3×3 weight matrix, and b_z is a bias term. The global multimodal embedding S_g is then fed into the channel attention module, which is compressed into two one-dimensional vectors by average pooling and maximum pooling, which are then passed through a shared Multilayer Perceptron (MLP) and finally normalized to the interval [0,1] by the sigmoid function to obtain the M_{CAM}:

$$M_{CAM} = \sigma(MLP(\eta(S_g)) + MLP(\gamma(S_g))) \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and η and γ represent average pooling and maximum pooling, respectively. Similarly, in the SAM, average pooling and maximum

pooling are again performed to aggregate the feature information and generate the 2D spatial attention map M_{SAM} using a convolutional layer of size 7×7:

$$M_{SAM} = \sigma(f^{7\times7}([\eta(M_{CAM}); \gamma(M_{CAM})])) \tag{4}$$

where $f^{7\times7}$ represents a convolutional layer of size 7×7, and η and γ represent average pooling and maximum pooling, respectively. Accordingly, the augmented feature S_g adjusted by CAM and SAM weighting is denoted as follows:

$$\bar{S}_g = M_{SAM} \otimes M_{CAM} \tag{5}$$

where $\otimes$ denotes the elemental multiplication. The dimensions are then restored to the original modal features using a fully connected layer:

$$R_s = W_s \bar{S}_g + b_s \tag{6}$$

where W_s and b_s represent the fusion weight matrices and bias terms of the fully linked network. Finally, the original input features are recalibrated using a gating mechanism:

$$\tilde{F}_s^i = 2 \times \sigma(R_s) \otimes F_s^i \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $f^{7\times7}$ denotes the elemental multiplication, and the coefficient 2 in Equation (7) serves as an amplification factor to further enhance the impact of the important features and ensure that the important features can receive more attention during the feature importance adjustment process. Overall, the textual, acoustic, and visual features after UFEM augmentation can be described as follows:

$$\tilde{F}_s^i = UFEM(F_s^i; \theta^{UFEM}) \in \mathbb{R}^{l_s \times d_s} \tag{8}$$

where θ^{UFEM} represents all the learnable parameters in the UFEM.

Sparse Phased Transformer: In the multimodal task, we use the Sparse Phased Transformer, SPT [13], architecture to extract the respective final feature representations from the data of different modalities. For any unimodal feature $\tilde{F}_s^i$, the final feature representation obtained after applying the SPT can be expressed as follows:

$$\tilde{F}_s^* = SPT(\widetilde{F}_s^i; \theta^{spt}) \tag{9}$$

where θ^{spt} is the learnable parameter of the SPT, and $s \in \{t, a, v\}$. To obtain the fused feature representation, we first concatenate each unimodal feature representation and then project each one into a lower-dimensional feature space $\mathbb{R}^{d_c}$. This process can be specifically expressed through linear transformation:

$$F_m^* = ReLU(W_1^m[\tilde{F}_t^*; \tilde{F}_a^*; \tilde{F}_v^*] + b_1^m) \tag{10}$$

where $\tilde{F}_t^*; \tilde{F}_a^*; \tilde{F}_v^*$ denote the final eigenvectors of the text, audio, and visual modalities, respectively, and W_1^m and b_1^m are the corresponding fusion weight matrices and bias terms. Finally, sentiment prediction based on the fused multimodal feature vectors is implemented:

$$\hat{y}_m = W_2^m F_m^* + b_2^m \tag{11}$$

where F_m^* is the fused multimodal eigenvector, W_2^m and b_2^m represent the weight matrix and bias term of the sentiment prediction output layer, respectively, and $\hat{y}_m$ is the predicted sentiment label.

3.4. Unimodal Task

In the three unimodal tasks, we adopt the same modal characterization approach as the multimodal task, thus mapping each feature representation to the common semantic feature space $\mathbb{R}^{d_c}$ as follows:

$$F_s^* = ReLU(W_1^s \tilde{F}_s^* + b_1^s) \tag{12}$$

where $s \in \{t, a, v\}$. Next, the feature representations for each modality are further processed through their respective independent fully connected layer networks to obtain the corresponding sentiment prediction output for each modality:

$$\hat{y}_s = W_2^s F_s^* + b_2^s \tag{13}$$

In order to facilitate the training process of the unimodal task, we have developed a novel ULGM, which is capable of generating unimodal labels. A detailed description of the specific architecture of the ULGM and its working principle will be provided in Section 3.6. The ULGM is calculated as follows:

$$y_s = ULGM(y_m, F_m^*, F_s^*, \theta^{ULGM}) \tag{14}$$

where y_m stands for multimodal labels, and θ^{ULGM} stands for ULGM learnable parameters. Finally, we adopted a joint learning strategy that combines the manually annotated multimodal label y_m and the automatically generated single modal label y_s to jointly train a multimodal task and three unimodal subtasks that are only relevant during the training phase. It is important to emphasize that these unimodal tasks only exist during the training period. Consequently, we utilize $\hat{y}_m$ as the ultimate result.

3.5. Hybrid Contrastive Learning

Unsupervised Contrastive Learning: Although the SPT successfully improves the expressiveness of fused features, it does not deeply explore the intrinsic connections between unimodal features F_s^i and fused features F_m^*. Therefore, we use Unsupervised Contrastive Learning (UCL) with the aim of strengthening these connections and further optimizing the quality of fusion features. The goal of our design is to maximize the mutual information between the fused features and the inputs of each unimodal modality, which is optimized through repeated iterative optimization; thus, the network can effectively transition from each independent modality to the fusion features. Given that the current Self-HCL has obtained the multimodal fusion result F_m^* via the SPT network, an effective mapping from the fusion feature F_m^* back to each unimodal input F_s^i has not yet been established. Therefore, we follow the operation of [10] and adopt a strategy to measure the correlation between them using a function $Corr(\cdot)$ with normalized prediction vectors and true vectors, which is defined as follows:

$$\bar{G}_\varphi(F_m^*) = \frac{G_\varphi(F_m^*)}{||G_\varphi(F_m^*)||_2}, \bar{F}_s^i = \frac{F_s^i}{||F_s^i||_2} \tag{15}$$

$$Corr(F_s^i, F_m^*) = exp(\bar{F}_s^i(\bar{G}_\varphi(F_m^*))^T) \tag{16}$$

where G_φ is a neural network with parameter φ that generates the prediction of F_s^i from F_m^*, and $||\cdot||_2$ is the $L2$ normalization. The loss between individual modalities and fused features is computed by treating all other modal representations as negative samples in the same batch of samples:

$$\mathcal{L}_{F_m^*, F_s^i} = -\mathbb{E}_s \left[log \frac{Corr(F_m^*, F_s^i)}{\sum_j^N Corr(F_m^*, F_s^i)} \right] \tag{17}$$

where N is the number of samples in the batch, and $\mathcal{L}_{F_m^*,F_s^i}$ denotes the contrastive learning loss function between the two vectors F_m^* and F_s^i. Ultimately, the overall loss function of the UCL consists of the sum of the losses of the fused features F_m^* with respect to the textual, visual, and audio modalities:

$$\mathcal{L}_{UCL} = \mathcal{L}_{m,t} + \mathcal{L}_{m,a} + \mathcal{L}_{m,v} \tag{18}$$

where m represents the fusion feature F_m^*.

Supervised Contrastive Learning: By utilizing the label information to the fullest, Supervised Contrastive Learning (SCL) treats all samples in the collection with the same label as positive samples and those with different labels as negative samples, thus presuming that attention will be paid to specific key labels. In particular, when dealing with datasets such as CMU-MOSI and CMU-MOSEI, which are only labeled with multimodal labels but not unimodal labels, the SCL approach can skillfully utilize the label information to achieve efficient feature learning and expression enhancement. Specifically, the model first encodes the different modal features (e.g., text, audio, visual) corresponding to the samples within each batch into consistent high-dimensional vectors. Embeddings of similarly labeled samples will be close to each other during the comparison learning process, while dissimilarly labeled samples will be far away from each other. This facilitates Self-HCL to capture potential semantic associations between different modalities related to specific sentiment categories and to combine information from multiple modalities to accomplish effective sentiment recognition tasks despite the lack of unimodal fine-grained labeling. The SCL loss $\mathcal{L}_{SCL}$ is computed as follows:

$$Z = [F_t^i; F_a^i; F_v^i; F_m^*] \tag{19}$$

$$SIM(p,i) = log\frac{\exp\left((Z_i \cdot Z_p)/\tau\right)}{\sum_{a\in A(i)} \exp\left(Z_i \cdot Z_p/\tau\right)} \tag{20}$$

$$\mathcal{L}_{SCL} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} SIM(p,i) \tag{21}$$

where $Z \in \mathbb{R}^{L\times d}$, $i \in I = \{1,2,...,L\}$ denotes the index of a batch of samples, $\tau \in \mathbb{R}^+$ denotes the temperature coefficient used to control the distances between the samples, $P(i) = I_{j=i} - \{i\}$ denotes the samples that share the same sentiment category as i but exclude i itself, $P(i)$ denotes the number of samples, and $A(i) = I - \{i\}$ denotes the samples in a batch of samples other than itself.

3.6. ULGM

The objective of the ULGM is to generate labels for each unimodality by applying multimodal labels and modality representations. Our ULGM design has been extended and optimized based on the work of Hwang et al. [11], whose design concept is that the distance between two features in the common semantic feature space is proportional to the distance between the corresponding labels in the Label Space. Based on this concept, and combining the features of unsupervised contrastive learning, we propose the Unsupervised Contrastive Learning Space (UCL Space). In the UCL Space, we map the data of different modalities into a unified representation space. In this space, if data points have similar attributes, they tend to be close to each other and form tight clusters, thus reflecting the similarity between data points. In contrast, data points that belong to different categories or have significant differences will be mapped to the far end of the space, thus highlighting the differences between them. The architecture of these three feature spaces is illustrated in Figure 3. In summary, the ULGM scheme is based on two key assumptions and mechanisms:

(1) The Common Semantic Feature Space is consistent with Label Space: The distance $D^F_{m\to s}$ between the eigenvectors of Fusion feature F^*_m and the eigenvectors of the unimodal feature F^*_s should be proportional to the semantic or categorical distance $D^L_{m\to s}$ between the labels of the two modalities corresponding to the two modalities in the Label Space.

(2) The Common Semantic Feature Space is associated with the UCL Space: The distance $D^F_{m\to s}$ within the feature space matches the relative position $D^C_{m\to s}$ between the fusion feature F^*_m and unimodal feature F^*_s embodied in the unsupervised contrastive learning. In summary, the design philosophy of the ULGM can be summarized as follows:

$$D^F_{m\to s} \propto D^L_{m\to s}, D^F_{m\to s} \propto D^C_{m\to s} \tag{22}$$

where $s \in \{t, a, v\}$. The ULGM method proposed in this work determines the amount of deviation of a unimodal label y_s with respect to a multimodal label y_m by measuring the distance from the multimodal feature to each unimodal feature. In the process of calculating the deviation, we focus on two core elements: the magnitude and the direction.

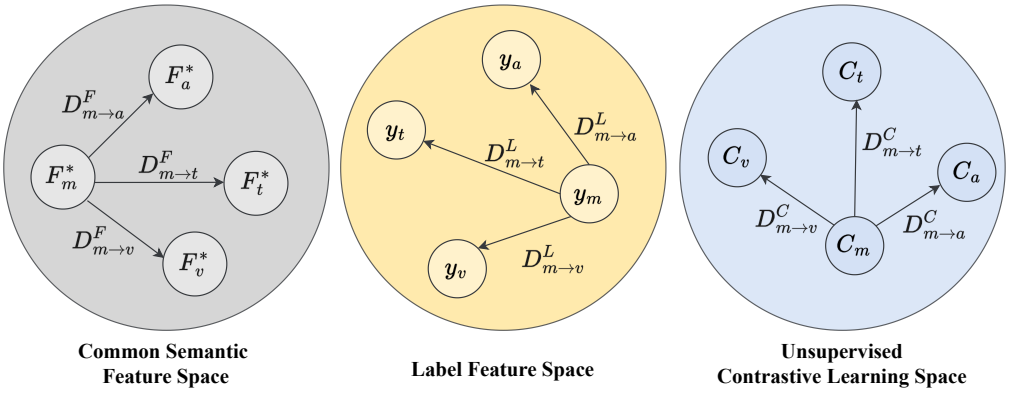


Figure 3. Schematic representation of the Common Semantic Feature Space, the Label Space, and the UCL Space.

Magnitude of Offset: To compute the offset, we argue that the greatest distance inside the common semantic feature space is proportional to the maximum distance within the Label Space. In the CMU-MOSI and CMU-MOSEI datasets, the multimodal labels vary from -3 to $+3$. This means that the distance between multimodal features with labels -3 (F^{*-3}_m) and $+3$ (F^{*+3}_m) must be the largest within the common same semantic feature space. Therefore, any $D^F_{m\to s}$ higher than the maximum distance is clipped to $D^F_{max} = ||\overline{F^{*+3}_m} - \overline{F^{*-3}_m}||$:

$$D^F_{m\to s} = \begin{cases} ||F^*_m - F^*_s||, & \text{if } \mathrm{D}^F_{m\to s} \le D^F_{max}, \\ D^F_{max}, & \text{otherwise,} \end{cases} \tag{23}$$

where $\overline{F^{*+3}_m}$ and $\overline{F^{*-3}_m}$ are the mean values of F^{*+3}_m and F^{*-3}_m, respectively, and $||\cdot||_2$ is the $L2$ normalization. Based on the concepts and points mentioned, we can consider the following relations to calculate the offset magnitude from multimodal to unimodal labels:

$$D^F_{m\to s}/D^F_{max} = D^L_{m\to s}/D^L_{-3\to+3} \tag{24}$$

$$D^L_{m\to s} = \frac{D^F_{m\to s}}{D^F_{max}} D^L_{-3\to+3} \tag{25}$$

Under the current conditions, the unimodal labels y_s can be estimated as follows:

$$y_s = y_m + D^L_{m\to s} \tag{26}$$

For the results of UCL, due to its wider range, it is necessary to define a maximum distance that is consistent with the previous setting. Therefore, we set $D^C_{max} = ||\overline{F^{*+3}_m} - \overline{F^{*-3}_m}||$. In order to establish the connection between $D^F_{m\to s}$, $D^C_{m\to s}$, and y_s, y_m, we consider the following two relations:

$$\frac{y_s}{y_m} \propto \frac{D^c_{m\to s}}{D^c_{max}} \Rightarrow \frac{y_s}{y_m} = \frac{D^c_{m\to s}}{D^c_{max}} \Rightarrow y_s = \frac{D^c_{m\to s}}{D^c_{max}} y_m \tag{27}$$

$$y_s - y_m \propto D^c_{m\to s} - D^c_{max} \Rightarrow y_s = D^c_{m\to s} - D^c_{max} + y_m \tag{28}$$

Combining the above relations, the unimodal label y_s in this condition is obtained using equal weight summation:

$$y_s = y_m + \varphi_{cm} \tag{29}$$

where $\varphi_{cm} = y_m(\frac{D^c_{m\to s} - D^c_{max}}{2D^c_{max}}) + \frac{D^c_{m\to s} - D^c_{max}}{2}$.

Direction of Offset: In order to determine the direction of the offset, the spatial location of the unimodal features relative to the multimodal features is first analyzed. This process first involves obtaining the average of the multimodal features with positive annotations $\overline{F^{*+}_m}$ and negative annotations $\overline{F^{*-}_m}$ as a reference datum. Then, with reference to this benchmark, the multimodal features and unimodal features are localized in the feature space, as shown in Figure 4. By calculating the $L2$ distances from various types of modal representations (e.g., $F^*_{x\in\{m,t,a,v\}}$) to $\overline{F^{*+}_m}$ and $\overline{F^{*-}_m}$, the directions of the offsets can be deduced and determined accordingly:

$$Direction = \begin{cases} +, & \text{if } \frac{D^p_s}{D^n_s} < \frac{D^p_m}{D^n_m}, \\ -, & \text{if } \frac{D^p_s}{D^n_s} > \frac{D^p_m}{D^n_m}, \\ 0, & \text{if } \frac{D^p_s}{D^n_s} = \frac{D^p_m}{D^n_m}. \end{cases} \tag{30}$$

where $D^p_s = ||F^*_s - \overline{F^{*+}_m}||$, $D^n_s = ||F^*_s - \overline{F^{*-}_m}||$, $D^p_m = ||F^*_s - \overline{F^{*+}_m}||$, $D^n_m = \|F^*_m - \overline{F^{*-}_m}\|$, and $\|\cdot\|$ are the $L2$ normalizations. Finally, the unimodal label y_s is obtained as follows:

$$y_s = \begin{cases} y_m + \alpha \times D^L_{m\to s} + \beta \times \varphi_{cm}, & \text{if direction is } +, \\ y_m - \alpha \times D^L_{m\to s} - \beta \times \varphi_{cm}, & \text{if direction is } -, \\ y_m, & \text{if direction is } 0. \end{cases} \tag{31}$$

where α and β represent the Label Space weight coefficients and the UCL Space weight coefficients, respectively.

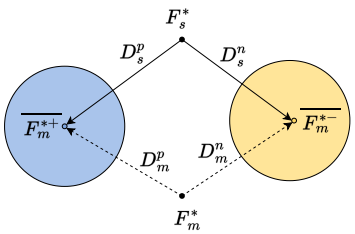


Figure 4. An illustration of the position of modality representations relative to the mean of multimodal representations with $\overline{F^{*+}_m}$ and $\overline{F^{*-}_m}$.

3.7. Objective Function for Training

We use the $\mathcal{L}_1$ loss as the main optimization objective of the model. In the unimodal task s, we use the difference between the automatically generated unimodal labels and the manually annotated multimodal labels as the weight of the loss function. This design means that the network will pay more attention to samples with large label differences, thereby improving the model's sensitivity to key differences. In addition, the unimodal task s provides an independent unimodal supervision signal and assists in multimodal task learning, thereby helping the model learn more discriminative modality-specific representations. The specific calculation formula is as follows:

$$\begin{aligned}\mathcal{L}_0 &= \mathcal{L}_1 + \frac{1}{N}\sum_{i}^{N}\sum_{s}^{\{t,a,v\}}(W_s^i \times |\hat{y}_s^i - y_s^i|) \\ &= \frac{1}{N}\sum_{i}^{N}(|\hat{y}_m^i - y_m^i|) + \frac{1}{N}\sum_{i}^{N}\sum_{s}^{\{t,a,v\}}(W_s^i \times |\hat{y}_s^i - y_s^i|) \\ &= \frac{1}{N}\sum_{i}^{N}(|\hat{y}_m^i - y_m^i| + \sum_{s}^{\{t,a,v\}} W_s^i \times |\hat{y}_s^i - y_s^i|)\end{aligned} \tag{32}$$

where N is the number of training samples. $W_s^i = \tanh(|y_s - y_m|)$ is the weight of the ith sample for the unimodal task s. The overall loss function $\mathcal{L}$ of Self-HCL combines the above components and is computed as follows:

$$\mathcal{L} = \lambda_0\mathcal{L}_0 + \lambda_1\mathcal{L}_{SCL} + \lambda_2\mathcal{L}_{UCL} \tag{33}$$

where λ_0 is the weight of the $\mathcal{L}_0$ loss, and λ_1 and λ_2 are the weights of $\mathcal{L}_{SCL}$ and $\mathcal{L}_{UCL}$, respectively, which are used to balance the contribution of different loss terms to model optimization.

4. Experimental Settings

4.1. Datasets

In this work, we conduct extensive experiments on two benchmark datasets in MSA. We give a brief introduction to each of them and summarize their basic statistics in Table 1.

CMU-MOSI: The CMU-MOSI dataset, introduced by [31], is widely acknowledged as a notable benchmark dataset for MSA. The dataset contains samples that have been annotated by human annotators with sentiment scores ranging from −3 (indicating strongly negative sentiment) to +3 (indicating very positive sentiment).

CMU-MOSEI: In contrast to CMU-MOSI, the CMU-MOSEI dataset [32] comprises a greater quantity of utterances, a more diverse sample of speakers, and a greater range of topics. In the same manner as MOSI, the CMU-MOSEI dataset is annotated with a sentiment score of −3 to +3 for each sample.

Table 1. Dataset statistics of CMU-MOSI and CMU-MOSEI.

Dataset	Train	Valid	Test	Total
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16,326	1871	4659	22,856

4.2. Baselines

In order to fully ensure the validity of Self-HCL, we provide a fair comparison between the baseline and state-of-the-art methods in the Multimodal Sentiment Analysis:

- **TFN** [6]: The Tensor Fusion Network (TFN) applies a subnetwork for modality embedding, along with tensor fusion, to understand both the intra- and intermodality dynamics.
- **LMF** [16]: Low-Rank Multimodal Fusion (LMF) carries out the fusion of multiple modalities by utilizing low-rank tensors, thus enhancing computational efficiency.

- **RAVEN** [33]: The Recurrent Attended Variation Embedding Network (RAVEN) captures the detailed structure of nonverbal subword sequences and adapts word representations in response to nonverbal signals.
- **MulT** [4]: The Multimodal Transformer (MulT) employs a crossmodal transformer with crossmodal attention to facilitate modality translation.
- **MISA** [5]: The Modality-Invariant and -Specific Representations (MISA) projects features into two separate spaces with specific constraints and performs fusion on these features.
- **MAG-BERT** [34]: The Multimodal Adaptation Gate for BERT (MAG-BERT) designs an alignment gate and inserts that into a vanilla BERT model to refine the fusion process.
- **Self-MM** [9]: Learning Modality-Specific Representations with Self-Supervised Multitask Learning (Self-MM) assigns each modality a unimodal training task with automatically generated labels, thus aiming to adjust the gradient backpropagation.
- **MMIM** [10]. Multimodal InfoMax (MMIM) uses the first implementation of the InfoMax principle on an MSA task, where the fusion representation is learned by maximizing its mutual information with unimodal representations.
- **SUGRM** [11]: The Self-Supervised Unimodal Label Generation Model (SUGRM) leverages recalibrated information to produce unimodal annotations by adaptively tuning features, thus postulating that the distance between two representations in a shared space should correspondingly reflect the distance between their associated labels in the label space.

4.3. Implementation Details

Experimental Details: Self-HCL was implemented on the Pytorch framework. For training the model, we used the Adam optimizer and implemented an early stopping strategy with eight cycles to monitor the performance of the model. To find the best combination of hyperparameters, we performed a stochastic search. Table 2 shows the detailed configuration of the CMU-MOSI and CMU-MOSEI datasets. All training and testing procedures were performed on a single NVIDIA GeForce RTX 3060 Ti GPU.

Evaluation Metrics: Following the previous works [9] , we report our experimental results in two forms: classification and regression. For classification, we report the weighted F1 score (F1-Score) and binary classification accuracy (Acc2). Specifically, for the CMU-MOSI and CMU-MOSEI datasets, we calculated the Acc-2 and F1-Score in two ways: negative/non-negative (nonexclude zero) and negative/positive (exclude zero). For the regression, we report the mean absolute error (MAE) and Pearson correlation (Corr). Except for the MAE, higher values denote better performance for all metrics.

Table 2. Main hyperparameters used in CMU-MOSI and CMU-MOSEI.

Hyperparameter	CMU-MOSI	CMU-MOSEI
Early Stop	8	8
Batch Size	32	32
LR for BERT	5×10^{-5}	5×10^{-5}
LR for Others	1×10^{-2}	1×10^{-3}
Encoder Layer	4	4
Num Heads	8	4
Output Dropout	0.3	0.1
Attn Dropout	0.3	0.1

5. Results and Analysis

5.1. Quantitative Results

The comparative results for the Multimodal Sentiment Analysis on the CMU-MOSI and CMU-MOSEI datasets are presented in Table 3. In this table, † means the results provided by MMIM [10], and ‡ is from SUGRM [11]. Models with * have been reproduced under the same conditions. Bold numbers indicate the best performance. Based on the

various types of datasets, they can be categorized as aligned or unaligned. Generally, models using aligned datasets will achieve superior performance [4]. In this work, we conducted experiments using unaligned datasets on our model. As described in Table 3, we achieved significant improvements in all the assessment metrics compared to the unaligned models (TFN and LMF). Even when compared with aligned models (RAVEN, MulT, MISA, and MAG-BERT), our approach achieved competitive results. In addition, we reproduced the three best baselines Self-MM, MMIM, and SUGRM under the same conditions. We found that our model outperformed them in most of the evaluations. Specifically, in the CMU-MOSI dataset, only MMIM outperformed our model in the evaluation metric of the MAE, which we analyze as a result of the fact that MMIM uses a historical data memory mechanism for entropy estimation, which ensures the stability and accuracy of the training process. And on the CMU-MOSEI dataset, our model successfully exceeded all baseline metrics and reached the optimal level.

Table 3. Experimental results on CMU-MOSI and CMU-MOSEI.

Model	CMU-MOSI				CMU-MOSEI				Data State
	Acc-2	F1-Score	MAE	Corr	Acc-2	F1-Score	MAE	Corr	
TFN †	- /80.8	- /80.7	0.970	0.698	- /82.5	- /82.1	0.593	0.700	Unaligned
LMF †	- /82.5	- /82.4	0.917	0.695	- /82.0	- /82.1	0.623	0.677	Unaligned
RAVEN ‡	-/78.0	-/76.6	0.915	0.691	-/79.1	-/79.5	0.614	0.662	Aligned
MulT †	81.5/84.1	80.6/83.9	0.861	0.711	- /82.5	- /82.3	0.580	0.703	Aligned
MISA †	80.79/82.10	80.77/82.03	0.804	0.764	82.59/84.23	82.67/83.97	0.568	0.724	Aligned
MAG-BERT ‡	82.5/84.0	82.4/84.0	0.778	0.766	81.3/84.8	81.7/84.7	0.567	0.742	Aligned
Self-MM	84.00/85.98	84.42/85.95	0.713	0.798	82.81/85.17	82.53/85.30	0.530	0.765	Unaligned
MMIM	84.14/86.06	84.00/85.98	0.700	0.800	82.24/85.97	82.66/85.94	0.526	0.772	Unaligned
SUGRM	84.4/86.3	84.3/86.3	0.703	0.800	83.7/84.4	83.6/84.0	0.544	0.748	Unaligned
Self-MM *	82.60/84.67	82.52/84.66	0.726	0.786	82.51/84.99	82.57/85.02	0.535	0.769	Unaligned
MMIM *	82.94/**84.91**	82.81/84.84	**0.707**	0.785	82.89/85.34	82.75/85.48	0.552	0.768	Unaligned
SUGRM *	82.36/83.99	82.35/84.04	0.727	0.776	82.85/83.81	82.94/83.83	0.542	0.742	Unaligned
Ours *	**83.14/84.91**	**83.17/84.96**	0.711	**0.788**	**83.12/85.91**	**83.19/85.93**	**0.531**	**0.775**	Unaligned

5.2. Ablation Study

Unimodal Task Analysis: To evaluate the contribution of unimodal tasks in Self-HCL, we conducted experiments to test the effects of different unimodal task combinations. As shown in Table 4, the overall performance of the model was improved after integrating unimodal tasks, and M, T, A, and V represent multimodal, text, audio and visual tasks, respectively. In the CMU-MOSI dataset, the model performance improved regardless of which modality task was added individually. In particular, the "M, A, T" and "M, V, T" combinations performed better than the "M, A, V" combination. A comparable phenomenon can be observed in the CMU-MOSEI dataset. To summarize, unimodal tasks have a positive effect on enhancing model performance. Specifically, text and audio modal tasks have been demonstrated to have a more significant influence on improving performance.

UFEM: To examine the efficiency of our proposed UFEM in improving unimodal features, we performed an ablation experiment using the baseline model SUGRM [11]. We made adjustments to SUGRM: we removed its modal feature calibration (MRM) component and implanted the UFEM for feature enhancement while keeping the other modules unchanged. The same adjustment was applied to the Self-HCL to compare the performance differences between the UFEM and MRM. Table 5 shows the performance comparison results of the two models on the unaligned datasets CMU-MOSI and CMU-MOSEI. The underlined numbers indicate improved performance compared to the baseline model. As can be seen in Table 5, when our model adopted MRM , its performance generally showed a downward trend. In contrast, when the SUGRM adopted our proposed UFEM, its overall

performance showed a significant improvement. This is attributed to the fact that the UFEM enhances the focus on key features and improves the expressiveness of the features, thus improving the performance of the model.

Table 4. Ablation study of unimodal task dominance using the unaligned datasets CMU-MOSI and CMU-MOSEI.

				CMU-MOSI				CMU-MOSEI
Task	**Acc-2**	**F1-Score**	**MAE**	**Corr**	**Acc-2**	**F1-Score**	**MAE**	**Corr**
M	81.78/83.73	81.80/83.91	0.729	0.775	82.19/84.15	82.70/84.42	0.548	0.757
M, T	82.13/83.93	82.07/83.99	0.737	0.783	82.40/84.70	82.42/84.15	0.538	0.758
M, A	82.20/84.06	82.19/84.13	0.748	0.772	82.70/84.77	82.90/84.60	0.543	0.762
M, V	81.47/84.23	82.51/84.06	0.742	0.769	82.23/83.52	82.36/83.73	0.546	0.751
M, A, V	82.99/84.77	82.55/84.56	0.722	0.782	82.23/84.23	82.68/85.29	0.544	0.761
M, A, T	83.02/84.92	**83.21**/84.95	0.728	0.783	**83.20**/85.43	83.07/85.51	0.543	0.762
M, V, T	82.92/**85.08**	82.72/84.86	0.718	0.775	82.23/85.23	82.68/**86.10**	**0.529**	0.757
M, T, A, V	**83.14**/84.91	83.17/**84.96**	**0.711**	**0.788**	83.12/**85.91**	**83.19**/85.93	0.531	**0.775**

Table 5. UFEM ablation study on the unaligned datasets CMU-MOSI and CMU-MOSEI.

					CMU-MOSI				CMU-MOSEI
Model	**Module**	**Acc-2**	**F1-Score**	**MAE**	**Corr**	**Acc-2**	**F1-Score**	**MAE**	**Corr**
SUGRM	MRM	82.36/83.99	82.35/84.04	0.727	0.776	82.85/83.81	82.94/83.83	0.542	0.742
	UFEM	82.47/84.23	82.45/84.27	0.723	0.779	83.02/84.23	83.11/84.43	0.538	0.756
Ours	MRM	82.84/84.52	82.93/84.47	0.718	0.780	82.94/85.63	82.96/85.71	0.536	0.762
	UFEM	**83.14/84.91**	**83.17/84.96**	**0.711**	**0.788**	**83.12/85.91**	**83.19/85.93**	**0.531**	**0.775**

HCL: In order to explore the impact of Hybrid Contrastive Learning (HCL) on our model performance, we conducted an ablation study on the unaligned datasets CMU-MOSI and CMU-MOSEI. Since HCL contains both Unsupervised Contrastive Learning (UCL) and Supervised Contrastive Learning (SCL) mechanisms, our ablation design was specified as follows:

- Employ w/o UCL: Remove only unsupervised contrastive learning from Self-HCL while leaving the rest unchanged.
- Employ w/o SCL: Remove only supervised contrastive learning from Self-HCL while keeping the remaining parts unaltered.

Table 6 shows the results of this ablation experiment. It is observed that when UCL was removed, the model showed a slight decrease in all the metrics, thus indicating that the UCL has a positive impact on improving the model's accuracy, F1-score, and Corr, as well as contributes to reducing the MAE. A similar trend can be observed when SCL was removed, thus confirming the effectiveness of HCL in enhancing the model in complex sentiment analysis tasks.

ULGM: The unique feature of our proposed ULGM is the introduction of a new unsupervised contrastive learning space, which is missing in the baseline model SUGRM [11]. Therefore, we did not directly apply the ULGM to the SUGRM, but we instead chose to perform ablation experiments within the Self-HCL framework. The specific settings are the following: $ULGM_{Ours}$ represents using our proposed ULGM in Self-HCL while ensuring that all other component configurations remain unchanged. For comparison, $ULGM_{SUGRM}$ represents the ULGM proposed using the SUGRM in Self-HCL while also keeping other components constant. Table 7 shows the results of the two processing methods on the unaligned CMU-MOSI and CMU-MOSEI datasets. We can observe from the table that when Self-HCL adopted the $ULGM_{SUGRM}$, various performance indicators of the model declined to varying degrees. This is because $ULGM_{SUGRM}$ faces challenges when dealing

with similarity modal features, while $ULGM_{Ours}$ takes full advantage of contrastive learning in mining feature differences by introducing a new UCL Space, thereby successfully solving the limitations of $ULGM_{SUGRM}$ and ultimately improving the overall performance of the model.

Table 6. Ablation study of HCL on the unaligned datasets CMU-MOSI and CMU-MOSEI.

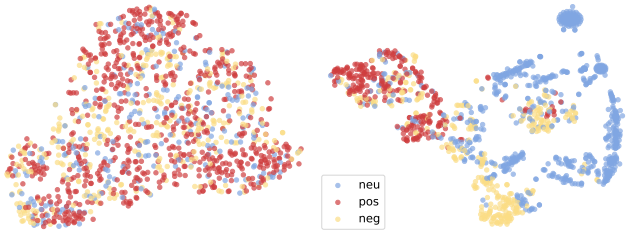
				CMU-MOSI				CMU-MOSEI
Model	**Acc-2**	**F1-Score**	**MAE**	**Corr**	**Acc-2**	**F1-Score**	**MAE**	**Corr**
w/o UCL	82.55/84.26	82.60/84.25	0.728	0.769	82.62/85.45	82.60/85.38	0.558	0.759
w/o SCL	82.78/84.23	82.86/84.57	0.722	0.773	82.87/85.68	82.86/85.68	0.546	0.762
Ours	**83.14/84.91**	**83.17/84.96**	**0.711**	**0.788**	**83.12/85.91**	**83.19/85.93**	**0.531**	**0.775**

Table 7. Ablation study of ULGM on the unaligned datasets CMU-MOSI and CMU-MOSEI.

				CMU-MOSI				CMU-MOSEI
Model	**Acc-2**	**F1-Score**	**MAE**	**Corr**	**Acc-2**	**F1-Score**	**MAE**	**Corr**
$ULGM_{SUGRM}$	82.49/84.30	82.58/84.33	0.727	0.768	82.50/85.47	82.66/85.58	0.552	0.760
$ULGM_{Ours}$	**83.14/84.91**	**83.17/84.96**	**0.711**	**0.788**	**83.12/85.91**	**83.19/85.93**	**0.531**	**0.775**

5.3. Case Study

HCL: To facilitate a qualitative examination of the Hybrid Contrastive Learning (HCL), we employed t-SNE [35] to visualize the preliminary distribution of some data and the hidden layer dynamics of the model subsequent to the application of HCL. As shown in Figure 5, the data without HCL processing had random distribution characteristics with no clear boundaries or clustering tendencies. In contrast, after applying HCL, the correlation between data points was optimized, the data points of the same category were aggregated to form a tight structure, and the separation between different categories was improved, thus showing stronger structure and recognizability. This shows that HCL plays a key role in improving model learning efficiency by strengthening feature fusion and contrastive learning, in addition to using multimodal label information to guide model training. Nevertheless, some data points may still be misclassified due to factors such as noise interference, modal mismatch, and sample complexity. Despite these problems, overall, HCL significantly improved the model's representation and classification performance for multimodal data. This finding prompts us to further optimize the learning strategy of the model to reduce misclassification.



(**a**) Before HCL (**b**) After HCL

Figure 5. T-SNE visualization of the embedding space.

ULGM: To evaluate the performance of the ULGM, we conducted experiments on the unaligned CMU-MOSI dataset. Figure 6 shows the trajectory of the unimodal labels, which gradually stabilized as the number of training iterations increased. After approximately 12 training epochs, the unimodal label distribution generated by the ULGM showed significant stability. Furthermore, to quantitatively evaluate the quality of the multimodal

labels generated by our model, we compared it with two baseline models: the Self-MM and SUGRM. Table 8 shows a detailed comparison of the fit between multimodal labels generated by different models and real labels. The results show that the multimodal labels generated by our proposed model fit the real labels more closely, which further proves the effectiveness and advancement of the ULGM.

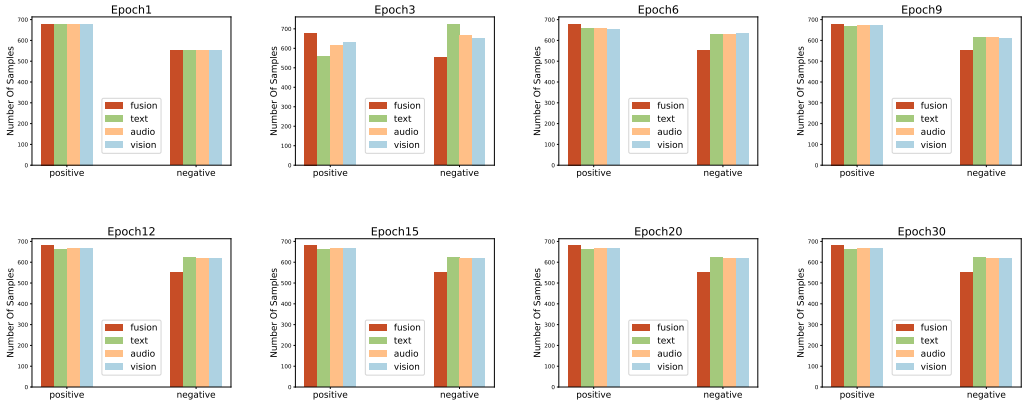


Figure 6. Visualization of the generated unimodal labels update process across epochs on the CMU-MOSI dataset.

Table 8. Case study for the Self-MM, SUGRM, and our model on the CMU-MOSI dataset.

Example	Annotation	Self-MM	SUGRM	Ours
Save your money wait till it comes out on rental. 	−2.0	−2.0	−1.9	−2.0
And I liked the first movie. I thought the first movie was really good. 	1.5	1.6	1.5	1.5
And I guess normally Shrek is for adults. 	0.0	0.1	0.1	0.0

6. Conclusions

In this work, we have presented a novel Multimodal Sentiment Analysis framework: Self-HCL. This framework optimizes the learning of unimodal feature representations in the absence of unimodal labeling by applying the Unimodal Feature Enhancement Module (UFEM), and it utilizes the Sparse Phased Transformer to capture and integrate the final feature representations for each modality. Furthermore, we implemented a Hybrid Contrastive Learning (HCL) strategy to enhance the representation of multimodal data and proposed a novel Unimodal Label Generation Module (ULGM) to generate stable

unimodal labels in a brief timeframe. Although Self-HCL introduces multiple optimization mechanisms, this may result in increased complexity and computational requirements for the model. However, we acknowledge that the introduction of multiple optimization mechanisms has increased the model's complexity and computational demands. This tradeoff between performance and computational efficiency is a critical consideration, especially in resource-constrained environments.

In light of these findings, we have identified avenues for future research. The primary focus will be on simplifying the model's architecture while striving to maintain or enhance its performance. This endeavor will involve exploring more lightweight components and algorithms that can offer comparable or superior results with reduced computational overhead. Moreover, we will delve deeper into the analysis of the results obtained, thus examining the impact of each component of Self-HCL on the overall performance. This comprehensive evaluation will provide valuable insights into the strengths and limitations of our framework, thus guiding further refinements and optimizations. Finally, we are committed to extending the applicability of Self-HCL to diverse domains and datasets, thus ensuring its robustness and versatility in real-world scenarios. By doing so, we aim to contribute to the broader field of sentiment analysis and pave the way for more sophisticated and efficient multimodal frameworks.

Author Contributions: Conceptualization, Y.F.; Funding acquisition, Y.F.; Investigation, Y.F. and J.F.; Methodology, Y.F. and J.F.; Project administration, Y.F.; Software, J.F.; Supervision, H.X.; Validation, J.F.; Visualization, J.F.; Writing—original draft, J.F.; Writing—review and editing, Y.F., J.F., H.X. and Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Chongqing Basic Research and Frontier Exploration Project (Chongqing Natural Science Foundation) [grant number: CSTB2022NSCQ-MSX0918], the Humanities and Social Sciences Project of Chongqing Education Commission [grant number: 23SKGH252] and the Chongqing University of Technology Graduate Education High-Quality Development Action Plan Funding Results [grant number: gzlcx20242041].

Data Availability Statement: This study utilized publicly available datasets from references [31,32].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
2. Grossberg, S. Recurrent neural networks. *Scholarpedia* **2013**, *8*, 1888. [CrossRef]
3. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
4. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019 ; Volume 2019, p. 6558.
5. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
6. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017.
7. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
8. Poria, S.; Hazarika, D.; Majumder, N.; Mihalcea, R. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.* **2020**, *14*, 108–132. [CrossRef]
9. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 10790–10797.
10. Han, W.; Chen, H.; Poria, S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv* **2021**, arXiv:2109.00412.
11. Hwang, Y.; Kim, J.H. Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, 2–6 May 2023; pp. 35–46.

12. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
13. Cheng, J.; Fostiropoulos, I.; Boehm, B.; Soleymani, M. Multimodal phased transformer for sentiment analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2447–2458.
14. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual information neural estimation. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
15. Kaur, R.; Kautish, S. Multimodal sentiment analysis: A survey and comparison. In *Research Anthology on Implementing Sentiment Analysis across Multiple Disciplines*; IGI Global: Hershey, PA, USA, 2022; pp. 1846–1870.
16. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
17. Liang, P.P.; Liu, Z.; Zadeh, A.; Morency, L.P. Multimodal language analysis with recurrent multistage fusion. *arXiv* **2018**, arXiv:1808.03920.
18. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6892–6899.
19. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8992–8999.
20. Zhang, Y.; Yang, Q. A survey on multi-task learning. *IEEE Trans. Knowl. Data Eng.* **2021**, *34*, 5586–5609. [CrossRef]
21. Yang, B.; Wu, L.; Zhu, J.; Shao, B.; Lin, X.; Liu, T.Y. Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2015–2024. [CrossRef]
22. Chauhan, D.S.; Dhanush, S.; Ekbal, A.; Bhattacharyya, P. Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4351–4360.
23. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
24. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
25. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
26. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
27. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
28. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; IEEE: Piscataway, NJ, USA , 2014; pp. 960–964.
29. iMotions, A. Facet iMotions Biometric Research Platform, 2013. Available online: https://imotions.com/products/imotions-lab/modules/fea-facial-expression-analysis/ (accessed on 16 July 2024).
30. Rakhlin, A. Convolutional neural networks for sentence classification. *GitHub* **2016**, *6*, 25.
31. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv***2016**, arXiv:1606.06259.
32. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. 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, Melbourne, Australia, 15–20 July 2018; Volume 1, pp. 2236–2246.
33. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7216–7223.
34. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating multimodal information in large pretrained transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Volume 2020, p. 2359.
35. Hinton, G.E.; Roweis, S. Stochastic neighbor embedding. *Adv. Neural Inf. Process. Syst.* **2002**, *15*, 857–864. [CrossRef]