Sequential Fusion of Text-close and Text-far Representations for Multimodal Sentiment Analysis

Kaiwei Sun and Mi Tian

Key Laboratory of Data Engineering and Visual Computing
Chongqing University of Posts and Telecommunications, Chongqing, China
sunkw@cqupt.edu.cn
s220231086@stu.cqupt.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) aims to identify human attitudes from diverse modalities such as visual, audio and text modalities. Recent studies suggest that the text modality tends to be the most effective, which has encouraged models to consider text as its core modality. However, previous methods primarily concentrate on projecting modalities other than text into a space close to the text modality and learning an identical representation, which does not fully make use of the auxiliary information provided by audio and visual modalities. In this paper, we propose a framework, Sequential Fusion of Text-close and Text-far Representations (SFTTR), aiming to refine multimodal representations from multimodal data which should contain both representations close to and far from the text modality. Specifically, we employ contrastive learning to sufficiently explore the information similarities and differences between text and audio/visual modalities. Moreover, to fuse the extracted representations more effectively, we design a sequential cross-modal encoder to sequentially fuse representations that are close to and far from the text modality. Experiments on three public benchmark datasets, MOSI, MOSEI, and CH-SIMS, demonstrate the superiority of the proposed method over the state-of-the-arts[1].

1 Introduction

Sentiment analysis has made remarkable advancements from the traditional textual sentiment classification which primarily relies on language to the more intricate Multimodal Sentiment Analysis (MSA) models (Zeng et al., 2022). Multimodal data provides not only verbal information, such as textual features but also non-verbal information, including acoustic and visual features (Hu et al., 2022). For instance, without audio and visual modalities, it is difficult to recognize the sentiment of "Yeah, I think so". Thus, combining different modalities together may help the machine to make decisions from different perspectives, thereby achieving more accurate predictions (Ngiam et al., 2011).

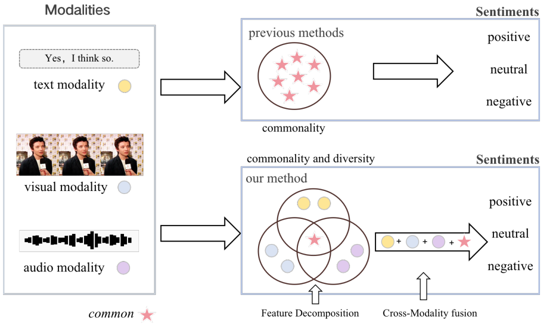


Figure 1: Previous methods versus our method.

However, multimodal learning (Baltrušaitis et al., 2018) processes heterogeneous information collected from multiple sources, which gives rise to two emergent issues: intra-modal representation and inter-modal fusion. Intra-modal representation learning mainly exploits consistency and complementarity of multiple modalities to bridge the gap between heterogeneous modalities. However, multiple works over-rely on text modality to improve the performance, so previous methods project audio and visual modalities into spaces close to the text modality to eliminate redundancy. But they neglect the fact that different modalities reveal distinctive characteristic of sentiment from different perspectives, and not take full advantage of the auxiliary information of visual and audio modalities. Accordingly, the key challenge of multimodal learning lies in two aspects: how to integrate commonality while preserving diversity of each individual modality and how to align different modality distributions interactively for inter-modal fusion (Zhang et al., 2022).

[1] The code is released at https://github.com/Mi7914/SFTTR

Proceedings of the 31st International Conference on Computational Linguistics, pages 40–49
January 19–24, 2025. ©2025 Association for Computational Linguistics

Based on the above motivation, we propose Sequential Fusion of Text-close and Text-far Representations for Multimodal Sentiment Analysis (SFTTR), Figure 1 illustrates the difference between previous methods and our proposed method. The main contributions are summarized as follows:

- A novel framework of Sequential Fusion of Text-close and Text-far Representations (SFTTR) is proposed. For intra-modal representation, we propose to decompose multiple modalities to two disjoint parts: Text-close and Text-far representations so as to extract similarities and differences between text and audio/visual modalities.

- For inter-modal fusion, we propose a novel sequential cross-modality encoder to sequentially fuse Text-close and Text-far representations.

- Experimental results on three public benchmark datasets, MOSI, MOSEI and CH-SIMS demonstrate that SFTTR achieves a new state-of-the-art performance.

2 Related Work

2.1 Multimodal Sentiment Analysis

There are many research directions in MSA, such as multimodal fusion (Yang et al., 2020), modal alignment (Tsai et al., 2019), context modeling (Mao et al., 2020) and so on. Early works of the first mainly operate geometric manipulation in the feature spaces (Zadeh et al., 2017). The recent works develop the reconstruction loss (Hazarika et al., 2020), or hierarchical mutual information maximization (Han et al., 2021) to optimize multimodal representation. For the modal alignment, Tsai et al. (2019) and Luo et al. (2021) leverage cross-modality and multi-scale modality representation to implement modal alignment, respectively. Lastly, studies of multimodal context integrate the unimodal context, in which (Chauhan et al., 2019) adapts context-aware attention, Ghosal et al. (2018) uses multi-modal attention, and Poria et al. (2017) proposes a recurrent model with multi-level multiple attentions to capture contextual information among utterances.

2.2 Contrastive Representation Learning

Contrastive learning has achieved great success in representation learning by contrasting positive pairs against negative pairs (Chen et al., 2020; Akbari et al., 2021; Hassani and Khasahmadi, 2020). Through a contrastive loss between augmented views of the same image sample, Chen et al. (2020) present a self-supervised framework, SimCLR, to learn visual representations. Khosla et al. (2020) extend self-supervised contrastive learning to the supervised setting, i.e., contrasting samples from different classes. Due to utilizing multimodal contrastive learning to train a Video-Audio-Text Transformer (VATT) for the alignment of video-text and video-audio pairs, Akbari et al. (2021) achieve state-of-the-art on various computer vision tasks, such as audio classification and visual action recognition. Hassani and Khasahmadi (2020) propose to learn node and graph level representations by contrasting encodings obtained from tdifferent structural views of graphs and achieve the state-of-the-art on various graph classification benchmarks. Yang et al. (2023) design a 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.

3 Method

3.1 Overall Architecture

The overall architecture of SFTTR is shown in Figure 2. As shown, SFTTR first extracts unified modality features from the input. After obtaining text, visual and audio features, we decompose each encoded modality into Text-close features (i.e., $C_T/C_V/C_A$ in Figure 2) and Text-far representations (i.e., $F_T/F_V/F_A$ in Figure 2) with different projectors. Finally we update the six decomposed features, fuse them in a sequential structure and gradually complement with each other.

3.2 Multimodal Input

Regarding the multimodal input, each sample consists of text (T), audio (A), and visual (V) sources. Referring to previous works, we use the [CLS] tag of BERT to encode text (i.e., T), and two separate transformer encoders to encode visual and audio modalities (i.e., V and A), respectively.

3.3 Feature Decomposition

A well-known fact in the MSA research is that the greater the difference between inter-modal representations, the better the complementarity of inter-modal fusion (Yu et al., 2020). Though

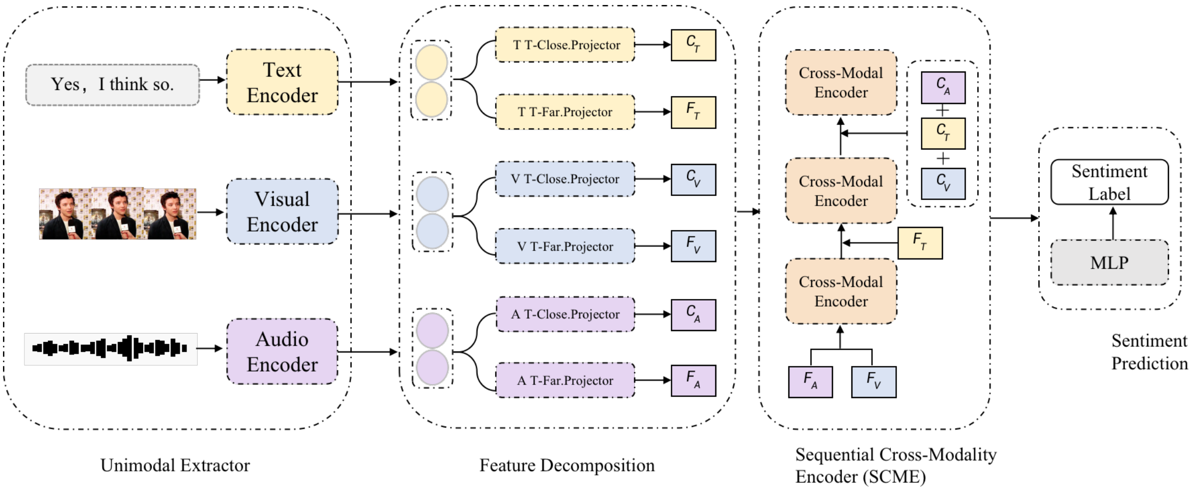


Figure 2: Overall structure of SFTTR.

the unimodal extractors capture long-term temporal context, they are unable to deal with feature redundancy due to modality gap (Zhang et al., 2022). To address this, we decompose each encoded modality into Text-close representations (i.e., C_T/C_V/C_A in Figure 2) and Text-far representations (i.e., F_T/F_V/F_A in Figure 2) with different projectors. Each projector consists of layer normalization, a linear layer with the Tanh activation, and a dropout layer. It inherently decomposes multiple modalities to two disjoint parts: Text-close and Text-far representations so as to extract information similarities and differences between text and audio/visual modalities.

Contrastive learning (CL) has gained significant advances in representation learning by viewing samples from various views (Gutmann and Hyvärinen, 2010; Khosla et al., 2020; Gao et al., 2021). The principle of contrastive learning is that an anchor and its positive sample should be pulled closer, while the anchor and negative samples should be pushed apart in feature space. In our work, we utilize contrastive learning to conduct modality decomposition. Previous works (Tsai et al., 2019; Yang et al., 2020) have demonstrated that textual modality is more indicative than the other modalities, to fully make use of the auxiliary information provided by audio and visual modalities, inspired by Yang et al. (2023), instead of treating all modalities equally as in other contrastive learning schemes, here we choose the text similarity feature C_T^i as an anchor, such that the visual and audio similarity features C_V^i and C_A^i are pushed closer to C_T^i, while in the meantime, the dissimilarity features in all modalities are pushed away from C_T^i. This allows the visual and audio similarity features to be drawn closer to the anchor, while simultaneously distancing the dissimilarity features in all modalities from it.

Specifically, denote the set of samples in a batch as $\mathcal{B}$, for each sample pair (i, j) in $\mathcal{B}$, we first calculate the cosine similarity score of them:

$$\mathrm{Cos}^{i,j} = \mathrm{sim}([T^i; V^i; A^i], [T^j; V^j; A^j]), \quad (1)$$

Subsequently, for each sample i, we sort samples with the same multimodal label y_m^i in ascending order of similarity scores to form the similar sample set S_0^i. In contrast, we sort samples that are different from y_m^i as the dissimilar sample set S_1^i. We randomly select two similar samples with high cosine similarity scores from S_0^i to form inter-sample positive pairs with sample i, which is denoted as $Neighbour^i$($\mathcal{N}^i$ for short). In the following, $\mathcal{N}^i$ will be used to refer to $Neighbour^i$. From the dissimilar sample set S_1^i, we select four samples to form inter-sample negative pairs. We denote them as $Outlier^i$($\mathcal{O}^i$ for short), where two samples have low cosine similarity scores and the other two have high scores. Choosing samples of high cosine similarity to form $\mathcal{O}^i$ can increase the difficulty of contrastive learning, prompting the model to better distinguish between similar yet different samples. This approach is particularly effective because it forces the model to focus on subtle differences, thereby enhancing its ability to discern nuanced features that are crucial for accurate classification.

Based on the samples obtained through these steps, the inter-sample positive/negative pairs for

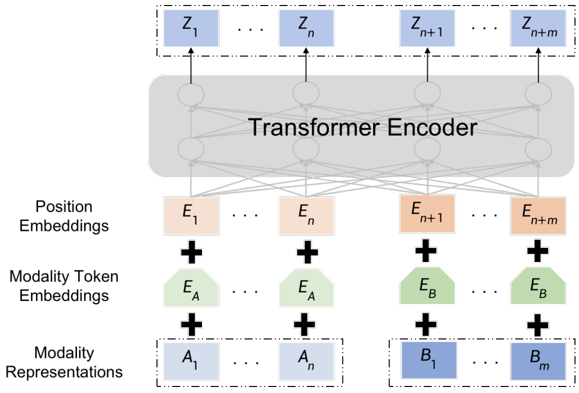


Figure 3: Structure of Cross-Modality Encoder (CME).

sample i are given by:

$$\begin{aligned}P^i_{inter} = &\{(C^i_T, C^j_T), (C^i_V, C^j_V), (C^i_A, C^j_A)| \\ &j \in \mathcal{N}^i\},\end{aligned} \tag{2}$$

$$\begin{aligned}N^i_{inter} = &\{(C^i_T, C^k_T), (C^i_V, C^k_V), (C^i_A, C^k_A)| \\ &k \in \mathcal{O}^i\},\end{aligned} \tag{3}$$

P^i_{intra} and N^i_{intra} are given by:

$$\begin{aligned}P^i_{intra} = &\{(C^i_T, C^i_V), (C^i_V, C^i_A)\} \cup \{(C^j_T, C^j_V), \\ &(C^j_T, C^j_A)|j \in \mathcal{N}^i \cup \mathcal{O}^i\},\end{aligned} \tag{4}$$

$$\begin{aligned}N^i_{intra} = &\{(C^i_T, F^i_T), (C^i_T, F^i_V), (C^i_T, F^i_A)\} \cup \\ &\{(C^j_T, F^j_T), (C^j_T, F^j_V), (C^j_T, F^j_A)| \\ &j \in \mathcal{N}^i \cup \mathcal{O}^i\},\end{aligned} \tag{5}$$

The set P^i and N^i are given by:

$$P^i = P^i_{intra} \cup P^i_{inter}, \tag{6}$$

$$N^i = N^i_{intra} \cup N^i_{inter}, \tag{7}$$

To simultaneously perform modality representation learning and decomposition, we use NT-Xent contrastive loss framework (Chen et al., 2020) to calculate the loss for sample i as follows:

$$l^i_{cl} = \sum_{(a,p)\in P^i} -log \frac{exp(sim(a,p)/\tau)}{\sum_{(a,k)\in(N^i \bigcup P^i)} exp(sim(a,k)/\tau)}, \tag{8}$$

where (a, p) and (a, k) denote a pair of decomposed feature vectors either within a sample or across different samples.

3.4 Multimodal Fusion and Output

3.4.1 Sequential Cross-Modality Fusion

The Text-close and Text-far representations that we obtain contain information similarities and differences between text and audio/visual modalities, while few or no information concerning modality interactions. Therefore, we need to fuse them into a joint representation for sentiment analysis. However, simply concatenating them together ignores modality interactions, which might introduce redundant information and lead to suboptimal problem (Zhang et al., 2018). Inspired by Zhang et al. (2022), we propose a novel Sequential Cross-Modality Encoder to exploit modality interactions. On the one hand, the feature distribution of various modalities varies due to heterogeneity, presenting a significant challenge to multimodal fusion. On the other hand, to preserve the temporal information of the two modalities, we augment them with positional embeddings. To bridge the large gap of the statistical properties between two modalities, we add two modality token embeddings to capture statistical regularities. The structure of the Cross-Modality Encoder (CME) is depicted in Figure 3, where the sum of modality representations, position embeddings, and modality token embeddings is feed into a Transformer Encoder, outputting the joint representation of modality A and B. While there have been previous studies on cross-modality fusion methods, our modules are more straightforward and easier to train without the need for additional hyper-parameter tuning. Cross-Modality Encoder can be written as $Z = \text{CME}(A, B)$.

Besides, on the one hand, audio and visual modalities tend to have stronger correlations in sentiment expression. On the other hand, they may include emotional information different from text modality. Therefore, combining these two modalities first may capture more emotional information (Wang et al., 2021), which is rarely considered in existing fusion methods. To remedy the deficiency, we devise Sequential Cross-Modality Encoder (SCME) to exploit interactions across modalities. Text-far representations $F_{\{V,A,T\}}$ and Text-close representations $C_{\{V,A,T\}}$ are fused in a sequential structure and gradually complement with each other. It is worth noting that, for the sake of brevity, we have provided simplified representations for the fusion of text-close representations in formulas and figures. Under this design, each pair of modalities interacts and correlates valuable information step by step, thus obtaining a joint multimodal representation M for final sentiment analysis. The sequential structure of multimodal fusion is given by:

$$Z_{VA} = \mathrm{CME}(F_V, F_A), \tag{9}$$

$$Z_{VAT} = \mathrm{CME}(Z_{VA}, F_T), \tag{10}$$

$$M = \mathrm{CME}(Z_{VAT}, C_V + C_A + C_T), \tag{11}$$

3.4.2 Overall Learning Objectives

After the multimodal fusion, we use a multilayer perceptron (MLP) with the ReLU activation function as the classifier to get the final predictive result. This choice is primarily driven by MLP's ability to capture complex nonlinear relationships within the input data, which is particularly crucial for sentiment analysis tasks, where the underlying patterns and nuances in the data are often highly non-linear and multifaceted. We use the joint multimodal representation M as the input to the classifier. Denote the set of samples in a batch as $\mathcal{B}$. For a given sample $i \in \mathcal{B}$, let its prediction from the classifier be $\hat{y^i_m}$, we calculate the multimodal prediction loss by mean squared error:

$$\hat{y^i_m} = \mathrm{MLP}(M), \tag{12}$$

$$L_{pred} = \frac{1}{n}\sum_{i=1}^{n}(y^i_m - \hat{y^i_m})^2, \tag{13}$$

where n is the number of samples in a batch and y^i_m is the multimodal label.

In addition, for each sample i, we also feed the 6 decomposed features $[C^i_T, C^i_V, C^i_A, F^i_T, F^i_V, F^i_A]$ into MLP classifier separately to get the 6 predictions denoted by the vector $\hat{u^i}$. Specifically, we compute the unimodal prediction loss by:

$$\hat{u^i} = \mathrm{MLP}([C^i_T, C^i_V, C^i_A, F^i_T, F^i_V, F^i_A]), \tag{14}$$

$$u^i = [y^i_m, y^i_m, y^i_m, y^i_T, y^i_V, y^i_A], \tag{15}$$

$$L_{uni} = \frac{1}{n}\sum_{i=1}^{n} \parallel u^i - \hat{u^i} \parallel^2, \tag{16}$$

where the vector $u^i = [y^i_m, y^i_m, y^i_m, y^i_T, y^i_V, y^i_A]$ represents the ground-truth labels for unimodal prediction. In other words, each decomposed feature is regularized to perform prediction individually.

Note that the Text-close features C^i_T, C^i_V, C^i_A are mapped through the MLP to predict the multimodal label y^i_m, whereas the Text-far features F^i_T, F^i_V, F^i_A are mapped through the MLP to predict modality-specific labels y^i_T, y^i_V, y^i_A(if available). Different from previous works, when modality-specific labels are not available, the unimodal prediction loss will no longer be considered. The rationale behind this design is that this may cause additional noise. It is worth mentioning that in the dataset we used, only the CH-SIMS include the modality-specific labels.

The contrastive loss and the overall loss function can be formulated as follows:

$$L_{cl} = \frac{1}{n}\sum_{i=1}^{n} l^i_{cl}, \tag{17}$$

$$L_{all} = L_{pred} + \beta_{uni}L_{uni} + \beta_{cl}L_{cl}, \tag{18}$$

where L_{pred} is the multimodal prediction loss, L_{uni} represents the unimodal prediction loss and L_{cl} represents the contrastive loss. β_{uni} and β_{cl} are hyper-parameters that balance the contribution of each regularization component to the overall loss L_{all}.

4 Experiments

4.1 Datasets

We conducted extensive experiments on three popular trimodal datasets (i.e., MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018b), and CH-SIMS (Yu et al., 2020)).

MOSI. The dataset comprises 2,199 multimodal samples encompassing visual, audio, and language modalities. Specifically, the training set consists of 1,284 samples, the validation set contains 229 samples, and the test set encompasses 686 samples. Each individual sample is assigned a sentiment score ranging from -3 (indicating strongly negative) to 3 (indicating strongly positive).

MOSEI. The dataset comprises 22,856 video clips collected from YouTube with a diverse factors (e.g., spontaneous expressions, head poses, occlusions, illuminations). This dataset has been categorized into 16,326 training instances, 1,871 validation instances, and 4,659 test instances. Each instance is meticulously labeled with a sentiment score ranging from -3 to 3. And the sentiment scores from -3 to 3 indicate most negative to most positive.

CH-SIMS. It is a Chinese multimodal sentiment dataset that comprises 2,281 video clips collected from various sources, such as different movies and TV serials with spontaneous expressions, various

	CH-SIMS					
Model	Acc-5 ↑	Acc-3 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
LF-DNN	41.62	66.91	78.87	79.87	0.420	0.612
MFN	39.47	65.73	77.9	77.88	0.435	0.582
LMF	40.53	64.68	77.77	77.88	0.441	0.576
TFN	39.30	65.12	78.38	78.62	0.432	0.591
MulT	37.94	64.77	78.56	79.66	0.453	0.561
MISA	-	-	76.54	76.59	0.447	0.563
MAG-BERT	-	-	74.44	71.75	0.492	0.399
Self-MM	41.53	65.47	80.04	80.44	0.425	0.595
ALMT	45.73	68.93	81.19	81.57	0.404	0.619
SFTTR	**47.48**	**70.24**	**81.62**	**81.66**	**0.368**	**0.6815**

Table 1: Results on CH-SIMS.

head poses, etc. It is divided into 1,368 training samples, 456 validation samples, and 457 test samples. Each sample is manually annotated with a sentiment score from -1 (strongly negative) to 1 (strongly positive).

4.2 Experimental Settings

We employ transformer encoders as our Vision Encoder and Audio Encoder. Specifically, for layer number in Transformer Encoder, we use two single-layer transformer encoders (Vaswani et al., 2017) on MOSI and CH-SIMS 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 of 0.0001 and batch size of 128. In the multimodal stage, we train SFTTR for MSA with the encoders obtained above. When modality-specific labels are available, we set the loss ratio to β_{cl} = 0.1 and β_{uni} = 0.01. While when they are not available, the loss ratio is set to be β_{cl} = 0.1 and β_{uni} = 0. For MOSI and CH-SIMS, we train SFTTR with the learning rate equals 0.0001 for 50 epochs. The batch size is set to 16 for MOSI and 32 for CH-SIMS. For MOSEI, we train the model for 25 epochs with a batch size of 4. The learning rate is set to 0.00005. All experiments were running with a single NVIDIA RTX 6000 GPU.

4.3 Evaluation Criteria

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. For classification, we report the multiclass 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 (Acc-2), 5-class prediction (Acc-5), 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 indicate better performance for all metrics.

4.4 Baselines

To comprehensively validate the performance of our SFTTR, we make a fair comparison with the several advanced and state-of-the-art methods, they can be grouped into 1) early multimodal fusion methods like Tensor Fusion Network **TFN** (Zadeh et al., 2017), Memory fusion network **MFN** (Zadeh et al., 2018a), Low-rank Multimodal Fusion **LMF** (Liu et al., 2018), and 2) methods that fuse multimodality through modeling modality interaction, such as Multimodal Transformer **MulT** (Tsai et al., 2019), **ALMT**(Zhang et al., 2023) which learns representation from other features under the guidance of language features and **PMR** (Lv et al., 2021) exchanges information with each modality by introducing a message hub, and 3) the methods focusing on the consistency and the difference of modality, in which **MISA**(Hazarika et al., 2020) controls the modal representation space, **Self-MM** (Yu et al., 2021) and **LF-DNN**(Yu et al., 2020) learns from unimodal representation using multi-task learning, **MAG-BERT** (Rahman et al., 2020) designs a fusion gate, **FDMER** (Yang et al., 2022) proposes a feature disentangled method to deal with modality heterogeneity by learning two distinct representations and **PS-Mixer** (Lin et al., 2023) realize better communication between different modal data.

4.5 Performance Comparison

Table 1 and Table 2 list the comparison results of our proposed method and state-of-the-art methods on CH-SIMS, MOSI and MOSEI respectively. The symbol ↑ denote higher values indicate better performance, the symbol ↓ is opposite. Besides, the best result is highlighted in bold.

It is worth noting that the scenarios in CH-SIMS are more complex than MOSI and MOSEI. Therefore, it is more challenging to model the multimodal data. However, as shown in the Table 1, our proposed method, SFTTR, consistently outperforms all other baselines on the CH-SIMS dataset on all metrics. For example, compared to ALMT,

	MOSI						MOSEI					
Model	Acc-7 ↑	Acc-5 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑	Acc-7 ↑	Acc-5 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
TFN	34.9	-	-/80.8	-/80.7	0.901	0.698	51.6	-	78.50/81.89	78.96/81.74	0.573	0.714
LF-DNN	34.52	-	77.52/78.63	77.46/78.63	0.955	0.658	50.83	-	80.60/82.74	80.85/82.52	0.58	0.709
LMF	33.2	-	-/82.5	-/82.4	0.917	0.695	51.59	-	80.54/83.48	80.94/83.36	0.576	0.717
MFN	34.1	-	77.4/-	77.3/-	0.965	0.632	51.34	-	78.94/82.86	79.55/82.85	0.573	0.718
MulT	40.0	-	-/83.0	-/82.8	0.871	0.698	52.84	-	81.15/84.63	81.56/84.52	0.559	0.733
MISA	42.3	-	81.8/83.4	81.7/83.6	0.783	0.776	52.2	-	**83.6/**85.5	**83.8/**85.3	0.555	0.756
MAG-BERT	41.43	-	82.13/83.54	81.12/83.58	0.790	0.766	50.41	-	79.86/**86.86**	80.47/83.88	0.583	0.741
PMR	40.6	-	-/83.6	-/83.4	-	-	52.5	-	-/83.3	-/82.6	-	-
FDMER	44.1	-	-/84.6	-/**84.7**	0.724	0.788	**54.1**	-	-/86.1	-/85.8	0.536	**0.773**
PS-Mixer	44.31	-	80.3/82.1	80.3/82.1	0.794	0.748	53.0	-	83.1/86.1	83.1/**86.1**	0.537	0.765
MulT*	-	42.68	-/-	-/-	-	-	-	54.18	-/-	-/-	-	-
MISA*	-	47.08	-/-	-/-	-	-	-	53.63	-/-	-/-	-	-
SFTTR	**46.5**	**52.62**	**82.94/84.6**	**82.92/**84.63	**0.709**	**0.795**	53.7	**55.48**	82.89/85.99	83.15/85.92	**0.536**	0.772

Table 2: Results on MOSI and MOSEI. * represents the result is from Mao et al. (2022).

it achieved relative improvements with 6.25% on Corr and 3.6% on MAE, respectively. Additionally, the proposed model demonstrates exceptional ability in multi-class classification, outperforming ALMT by 1.75% on Acc-5 and 1.31% on Acc-3. The superior classification performance demonstrates that our designed learning method is more effective than the compared methods. Furthermore, the significant improvement in Acc-2 and F1 further highlights the ability of our model to better understand the CH-SIMS dataset than the other baselines. Achieving such superior performance on CH-SIMS with more complex scenarios demonstrates SFTTR's ability to extract effective sentiment information from various scenarios.

As seen in the results in Table 2, on the MOSI dataset, our method outperforms all other baselines in all metrics except for the negative/positive (NP) setting F1 score. Furthermore, on the task of more difficult and finegrained sentiment classification (Acc-7), our model achieves a relative improvement of 2.4% compared to the secondbest result obtained by FDMR. For the MOSEI dataset, our model also surpass most of the baselines in all metrics. Specially, our method shows better performance in MAE and the negative/positive (NP) setting for F1 score. The Acc-7 and Corr are also better or comparable to most baselines.

4.6 Ablation Study and Analysis

To verify the effectiveness of each component of our SFTTR, in Table 3, we present the ablation result of the subtraction of each component on the CH-SIMS datasets. Among them, "-cl" denotes the removal of the contrastive learning method. "-uni" denotes the removal of the unimodal prediction component. "-fusion" represents the absence of the hierarchical cross-modality fusion. We observe that deactivating the hierarchical cross-modality fusion greatly decreases the performance, demonstrating it is effective. Moreover, after the removal of the the unimodal prediction task, the performance drops again, also supporting that the hierarchical cross-modality fusion and unimodal prediction task can effectively improve the SFTTR's ability to explore the sentiment information in each modality.

Model	Acc-5 ↑	Acc-3 ↑	Acc-2 ↑	F1 ↑	MAE ↓	Corr ↑
-cl	47.31	68.44	80.74	80.66	0.384	0.627
-uni	44.90	68.09	79.56	79.54	0.394	0.626
-fusion	44.41	68.44	80.39	80.12	0.402	0.631
SFTTR	**47.48**	**70.24**	**81.62**	**81.66**	**0.368**	**0.6815**

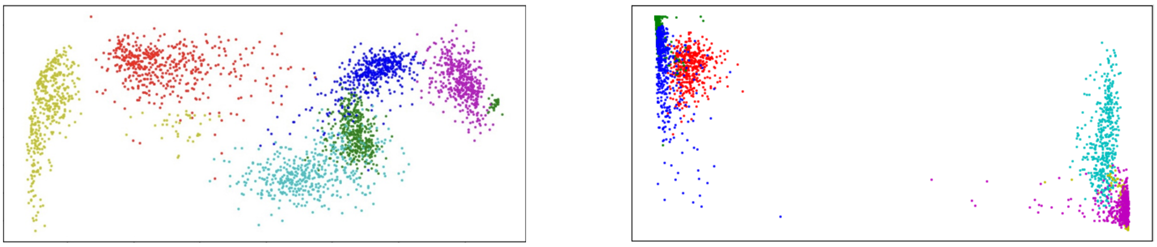
Table 3: Ablation study of SFTTR on CH-SIMS.

4.7 Visualization

Figure 4 shows the T-SNE visualization of all six Text-close and Text-far representations of all test samples on CH-SIMS, where (a) is the six decomposed features without SFTTR and (b) shows these features with SFTTR. From it, we can observe the distributions of Text-close features (i.e., in red, in green and in blue) become closer to each other while the Text-far features (i.e., in cyan, in yellow and in magenta) become further away from their corresponding Text-close features, proving the effectiveness of SFTTR to learn the similarities and differences between modalities.

5 Conclusion and Future Work

In this paper, we propose a novel method for multimodal sentiment analysis called SFTTR, consisting of uni-modal extractor, contrastive feature decomposition and sequential cross-modality fusion. These modules cooperate closely to capture the consistency and difference across modalities, fuse Text-far modality representations first and Text-close modality representations second, rather than merging the whole representations directly. While



(a)

(b)

Figure 4: T-SNE visualization comparison of all six Text-close and Text-far representations between: (a)case without SFTTR, and (b)case with SFTTR. The colors red, green, blue, cyan, yellow and magenta represent C_T, C_V, C_A, F_T, F_V, F_A respectively.

the proposed method achieved improved performance on several popular datasets, there are some limitations to consider. Firstly, as the number of training samples increases, our method may require more processing time. Additionally, the potential sentiment-irrelevant and conflicting information across modalities are not sufficiently alleviated. We plan to address this in future work, and the focus will be on designing a modules to capture the sentiment shifts for fine-grained sentiment prediction.

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*, 34:24206–24221.

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.

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5647–5657.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

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.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

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.

Han Lin, Pinglu Zhang, Jiading Ling, Zhenguo Yang, Lap Kei Lee, and Wenyin Liu. 2023. PS-Mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. *Information Processing & Management*, 60(2):103229.

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

Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. Scalevlad: Improving multimodal sentiment analysis via multiscale fusion of locally descriptors. *arXiv preprint arXiv:2112.01368*.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2554–2562.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-SENA: An integrated platform for multimodal sentiment analysis. *arXiv preprint arXiv:2203.12441*.

Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. 2020. Dialoguetrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1033–1038. IEEE.

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.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

C Wang, W Li, and Z Chen. 2021. Reserch of multimodal emotion recognition based on voice and video images. *Comput. Eng. Appl*, 57:163–170.

Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2020. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. *arXiv preprint arXiv:2010.11985*.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630.

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. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multiview 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: Cmumosei 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.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2022. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2924–2934.

Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang. 2018. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*.

Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9100–9108.