ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis

Yakun Yu [1], **Mingjun Zhao** [1], **Shi-ang Qi** [1], **Feiran Sun** [2], **Baoxun Wang** [2], **Weidong Guo** [2], **Xiaoli Wang** [2], **Lei Yang** [2], **Di Niu** [1]
[1]University of Alberta
[2]Platform and Content Group, Tencent
[1]{yakun2, zhao2, shiang, dniu}@ualberta.ca

Abstract

Multimodal Sentiment Analysis leverages multimodal signals to detect the sentiment of a speaker. Previous approaches concentrate on performing multimodal fusion and representation learning based on general knowledge obtained from pretrained models, which neglects the effect of domain-specific knowledge. In this paper, we propose Contrastive Knowledge Injection (ConKI) for multimodal sentiment analysis, where specific-knowledge representations for each modality can be learned together with general knowledge representations via knowledge injection based on an adapter architecture. In addition, ConKI uses a hierarchical contrastive learning procedure performed between knowledge types within every single modality, across modalities within each sample, and across samples to facilitate the effective learning of the proposed representations, hence improving multimodal sentiment predictions. The experiments on three popular multimodal sentiment analysis benchmarks show that ConKI outperforms all prior methods on a variety of performance metrics.

1 Introduction

Multimodal sentiment analysis (MSA) is the task of mining and comprehending the sentiments of online videos, which has many downstream applications, e.g., analyzing the overall opinion from customers about a product, gauging polling intentions from voters (Han et al., 2021; Melville et al., 2009), etc. Most existing MSA methods focus on developing fusion techniques between modalities. The easiest way is to simply concatenate text, video, and audio features as a fused vector for subsequent classification or regression. An alternative is to use outer-product, Recurrent Neural Networks (RNNs) or attention-based models to model multimodal interactions (Chen et al., 2017; Williams et al., 2018; Zadeh et al., 2017; Liu and Shen, 2018). More recently, MSA methods for learning effective multimodal representations have emerged constantly (Hazarika et al., 2020; Mai et al., 2021; Yu et al., 2021), ranging from decomposing the representation of each modality to introducing extra constraints in the learning objective.

Although the above methods have led to improvements in MSA performance, they focus on utilizing general knowledge obtained from pretrained models to encode modalities, which is inadequate to identify specific sentiments across modalities. One possibility to solve this issue is through knowledge injection which can generate specific knowledge to aid the general knowledge for further improving predictions. Many researchers have discovered that injecting knowledge from other sources such as linguistic knowledge, encyclopedia knowledge, and domain-specific knowledge can help enhance existing pretrained language models in terms of knowledge awareness and lead to improved performance on various downstream tasks (Wei et al., 2021; Lauscher et al., 2020; Wang et al., 2021a).

In this paper, we propose ConKI, a Contrastive Knowledge Injection framework, to learn both pan-knowledge representations and knowledge-specific representations to boost MSA performance. We argue that a unimodal representation can consist of a pan-knowledge representation (given by a pretrained model like BERT (Devlin et al., 2019)) and a knowledge-specific representation (injected from relevant external sources). Specifically, ConKI uses a pretrained BERT model to extract textual pan-knowledge representations and uses two randomly initialized transformer encoders to generate acoustic and visual pan-knowledge representations, respectively. In the meantime, it applies a knowledge injection model named adapter, onto each modality to yield knowledge-specific representations. Both pan- and specific-knowledge representations are fused first within each modality and then across modalities, before the fused features are used for sentiment prediction. We further propose a hierarchical contrastive learning procedure

Findings of the Association for Computational Linguistics: ACL 2023, pages 13610–13624

performed between knowledge types within every single modality, across modalities within each sample, and across samples, to facilitate the learning of these representations in ConKI.

The main contributions of this work can be summarized as follows:

- We propose ConKI, a Contrastive Knowledge Injection framework for multimodal sentiment analysis. ConKI aims to boost model performance through external knowledge injection from other datasets and hierarchical contrastive learning, which is proved better than simply fine-tuning with external datasets.
- We propose hierarchical contrastive learning that uses a unified contrastive loss to disentangle the pan-knowledge representations from the specific-knowledge representations since they belong to different knowledge domains and should complement each other.
- We conduct extensive experiments on three popular benchmark MSA datasets and attain results that are superior to the existing state-of-the-art MSA baselines on all metrics, demonstrating the effectiveness of the proposed methods in ConKI.

2 Related Work

In this section, we discuss related research in multimodal sentiment analysis, knowledge injection, and contrastive learning.

2.1 Multimodal Sentiment Analysis

Research on MSA mainly focuses on multimodal fusion and representation learning. For multimodal fusion, existing methods are typically divided into early fusion and late fusion techniques. Early fusion refers to joining multimodal inputs into a single feature before single-model encoding. For example, Williams et al. (2018) concatenate initial input features and then use LSTM to capture the temporal dependencies in the sequence. On the contrary, late fusion learns unimodal representations via separate models and fuses them in a later stage for inference. Zadeh et al. (2017) introduce a tensor fusion network that first encodes each modality with corresponding sub-networks and then models the unimodal, bimodal, and trimodal interactions by a three-fold Cartesian product. For representation learning methods, Hazarika et al. (2020) propose to project each modality into a modality-invariant and modality-specific representation. Different from the above work, we propose to decompose each modality into two representations based on knowledge types. Both representations can complement each other, leading to a richer unimodal representation.

2.2 Knowledge Injection

Injecting knowledge into pretrained language models (PLMs) has been proven to outperform vanilla pretrained models on various NLP tasks (Wei et al., 2021; Wang et al., 2021a; Tian et al., 2020; Ke et al., 2020; Lin et al., 2019; Wang et al., 2021b). Adapters are commonly used as a knowledge injection model plugged outside or inside of PLMs. For instance, Wang et al. (2021a) infuse factual knowledge from Wikidata (Vrandečić and Krötzsch, 2014) and linguistic knowledge from web text to RoBERTa (Liu et al., 2019) via two kinds of adapters. In this work, we build different adapters for different modalities, not limited to text, to learn specific multimodal knowledge from an external dataset for the downstream task. To the best of our knowledge, we are the first to explore knowledge injection in the multimodal domain.

2.3 Contrastive Learning

Contrastive learning (CL) aims to learn effective representations such that positive pairs of samples are close while negative pairs of samples are far apart (Liu et al., 2021; Li et al., 2020; Chen et al., 2020a; Khosla et al., 2020; He et al., 2020). Existing works can be divided into two categories: self-supervised CL (Akbari et al., 2021; Chen et al., 2020a,b; He et al., 2020; You et al., 2020; Tao et al., 2020) and supervised CL (Khosla et al., 2020; Mai et al., 2021). The difference between them is whether the label information is used to form positive/negative pairs. For example, Khosla et al. (2020) propose supervised CL to pull samples of the same class together and push samples from different classes away. In our work, we design contrastive pairs in finer granularity. That is, we consider contrasts between knowledge types, between modalities, and across samples.

3 Method

In this section, we explain the Contrastive Knowledge Injection framework (ConKI) in detail. The goal of ConKI is to generate pan- and specific-

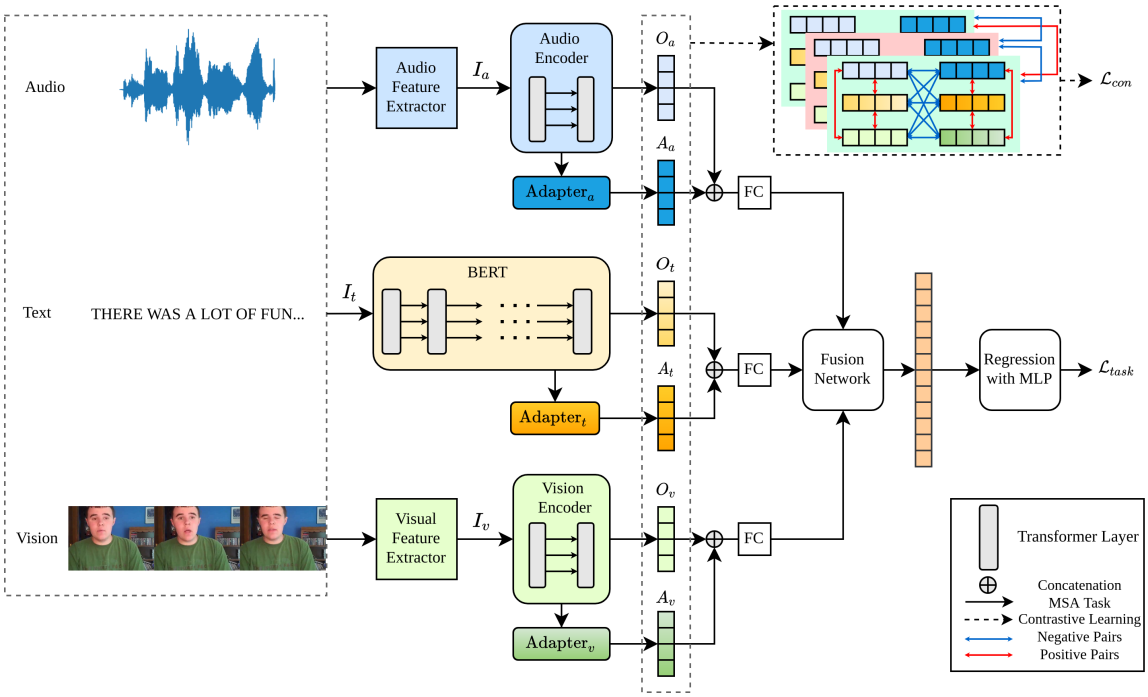

Figure 1: The overall architecture of ConKI. The solid and dashed arrows represent the procedure of the main MSA task and the hierarchical contrastive learning subtask, respectively. Inside the contrastive learning procedure, cyan and pink boxes illustrate samples that fall in different sentiment score intervals.

knowledge modality representations via knowledge injection and hierarchical contrastive learning. Knowledge injection intends to obtain knowledge-specific representations that could complement the pan-knowledge representations offered by pretrained models. Hierarchical contrastive learning further optimizes these knowledge-specific and pan-knowledge representations by considering contrasts between knowledge types, modalities, and samples.

3.1 Problem Definition

The task of multimodal sentiment analysis (MSA) is to detect sentiments in videos based on multimodal signals, including text (t), vision (v), and audio (a) modalities. These signals are represented as sequences of low-level features, i.e., $I_t \in \mathbb{R}^{l_t \times d_t}$, $I_v \in \mathbb{R}^{l_v \times d_v}$, and $I_a \in \mathbb{R}^{l_a \times d_a}$, respectively. Here $l_{m\in\{t,v,a\}}$ denotes the length of the sequence for each modality, while $d_{m\in\{t,v,a\}}$ denotes the corresponding feature vector dimension. The detail for acquiring these features is described in Appendix B. Given these sequences $I_{m\in\{t,v,a\}}$, the primary task is to make accurate predictions on the sentiment intensity by extracting and fusing higher-level multimodal information.

3.2 Overall Architecture

Figure 1 shows the overall architecture of ConKI. We first process raw multimodal input to low-level features $I_{m\in\{t,v,a\}}$ with their corresponding feature extractors and tokenizers. Then we encode I_m into knowledge-specific representations (i.e., A_m) generated by some adapters and pan-knowledge representations (i.e., O_m) generated by pretrained encoders. The text encoder is from publicly-available pretrained backbones like BERT (Devlin et al., 2019), and the vision/audio encoder is a designed model with random initialization since there is no suitable backbone that is pretrained by the above low-level features. After generating the knowledge-specific and pan-knowledge representations, ConKI is trained simultaneously with two different tasks on the downstream target dataset – the primary MSA regression task and the contrastive learning subtask.

For the MSA task, we concatenate the knowledge-specific representation and pan-knowledge representation of each modality before feeding them into a fully-connected (FC) layer for inner-modality fusion. We then design a fusion network that consists of a concatenation layer and a fusion module for multi-modality fusion, as shown in Figure 2. The fused representations are

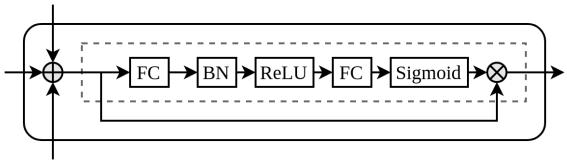

Figure 2: The Fusion Network. The fusion module marked in the dashed box is used to get the weighted fused embedding. $\otimes$ means element-wise multiplication.

passed into a multilayer perceptron (MLP) network to produce the sentiment predictions, $\hat{y}$.

For the subtask of hierarchical contrastive learning, we carefully construct the negative and positive sample pairs at the knowledge level, modality level, and sample level. The intuition of our pairing policy is as follows. We expect A_m and O_m to capture different knowledge, so we disentangle them and make them complement each other to get richer modality representations by knowledge-level contrasts. Since a video's sentiment is determined by all modalities, we learn the commonalities among the six representations by modality-level contrasts. Besides, videos that express close sentiments should share some correlations. We capture the correlations by sample-level contrasts to help further learn the commonalities among samples under close sentiments. By integrating these hierarchical contrasts, ConKI is able to catch full dynamics among representations which can significantly benefit the main MSA task.

3.3 Encoding with Knowledge Injection

We encode each modality into a pan-knowledge representation via the pretrained encoders and a knowledge-specific representation via the adapters.

Pan-knowledge representations. We use the pretrained BERT (Devlin et al., 2019) to encode the input sentence for the text modality. The pooled output vector in the last layer is extracted as the whole sentence representation O_t:

$$O_t, H_t = \text{Bert}(I_t; \theta_t^{\text{Bert}}) \ , \tag{1}$$

where H_t denotes the hidden states of all layers. For audio and vision modalities, we employ encoders of stacked transformer layers (Vaswani et al., 2017) to capture the temporal features O_m:

$$O_m, H_m = \text{Encoder}(I_m; \theta_m^{\text{encoder}}), \ m \in \{v, a\} \ . \tag{2}$$

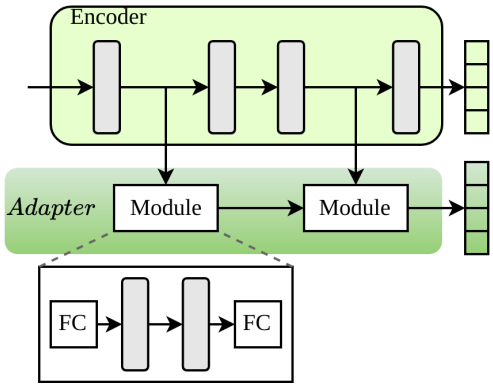

Figure 3: Adapter and its connection with the backbone encoder.

Here, O_t, O_a, and O_v are regarded as three pan-knowledge representations since they mainly contain general knowledge such as the generic facts encoded by BERT (Devlin et al., 2019) pretrained on big text data.

Knowledge-specific representations. We infuse specific domain knowledge from external multimodal sources through knowledge injection models (adapters). The adapter is commonly used in natural language processing (NLP) to enhance existing pretrained language models' knowledge awareness (Wei et al., 2021). The outputs of adapters are taken as knowledge-specific representations. Specifically, the adapter for each modality is plugged outside of the respective pretrained encoder, as shown in Figure 3. It consists of multiple modules with the same sandwich structure: two FC layers with two transformer layers in between. Each module can be inserted before any transformer layers of the backbone models (encoders), e.g., the second and fourth transformer layers in Figure 3. Therefore, each module takes the intermediate layers' hidden states of the pretrained encoder and the output of the previous adapter module as input. The output of the adapter is denoted as A_m, where

$$A_m = \text{Adapter}(H_m; \theta_m^{\text{adapter}}), \ m \in \{t, v, a\} \ . \tag{3}$$

With the objective of learning specific multimodal sentiment knowledge, we pretrain one adapter for each modality, i.e., Adapter_t, Adapter_a and Adapter_v, concurrently using an external dataset while keeping the pretrained encoders frozen. Since the external dataset we select is also from the multimodal sentiment domain, the pretraining task remains the MSA task. That is, we

Algorithm 1: Learning Procedure of ConKI

Stage 1: Adapter Pretraining

Input: External dataset $\mathcal{E}$, its corresponding features I_m and labels y,

Output: Pretrained adapters $\{\theta_m^{\text{adapter}} \mid m \in \{t, v, a\}\}$

for *each training epoch* **do**

  for *batch* $\{(I_t^i, I_v^i, I_a^i)\}_{i=1}^{|B|}$ *from* $\mathcal{E}$ **do**

1 Encode I_m^i to O_m^i and A_m^i via Eq. (1-3)

2 Inner-modality fusion: $F_m^i = FC([O_m^i; A_m^i])$, where $[\cdot;\cdot]$ denotes the concatenation of two vectors

3 Multi-modality fusion: $F^i = FN(F_t^i, F_v^i, F_a^i)$, where FN is the fusion network

4 Compute the predictions using $\hat{y^i} = MLP(F^i)$

5 Compute $\mathcal{L}_{task}$ via Eq. (4)

6 Update parameters except $\{\theta_t^{\text{Bert}}, \theta_m^{\text{encoder}} \mid m \in \{v, a\}\}$

  end

  Save $\{\theta_m^{\text{adapter}} \mid m \in \{t, v, a\}\}$ when reaching the best validation result

end

Stage 2: Downstream Fine-tuning

Input: Target dataset $\mathcal{D}$, its corresponding features I_m, its labels y, and the pretrained adapters

Output: Predictions $\hat{y}$

for *each training epoch* **do**

  for *batch* $\{(I_t^i, I_v^i, I_a^i)\}_{i=1}^{|B|}$ *from* $\mathcal{D}$ **do**

    Perform Steps $1-4$ in Stage 1

    Compute $\mathcal{L}$ via Eq. (5)

    Update parameters except $\{\theta_m^{\text{adapter}} \mid m \in \{t, v, a\}\}$

  end

end

pretrain adapter parts in Figure 1 with only the MSA task on the external dataset, then utilize the pretrained adapters to produce knowledge-specific representations A_m for the downstream target task that includes both the MSA task and the hierarchical contrastive learning subtask. Algorithm 1 summarizes this pretraining procedure of adapters.

3.4 Hierarchical Contrastive Learning

In our framework, we propose a hierarchical contrastive learning method to enhance the learned representations by considering the following four aspects in a batch B:

- For a single video sample i, all the modalities share common motives of the speaker that determine the overall sentiment. The pan-knowledge representations of different modalities are expected to represent similar meanings and thus need to be pulled closer to each other. And the same applies for knowledge-specific representations. This intuition leads to the construction of *intra-sample* positive pairs:

$$\mathcal{P}_1^i = \{(O_m^i, O_n^i), (A_m^i, A_n^i) \mid m, n \in \{t, v, a\} \ \& \ m \neq n \ \& \ i \in B\};$$

- The pan-knowledge representations and the knowledge-specific representations should be disentangled from each other since they belong to different knowledge domains and are designed to complement each other. This exists inside each sample (i and j represent the same sample) as well as across samples in the batch (i and j represent two different samples). Therefore, we can build the *inter-knowledge* negative pairs within a batch:

$$\mathcal{N}_1^i = \{(O_m^i, A_n^j) \mid m, n \in \{t, v, a\} \ \& \ i, j \in B\};$$

- For two arbitrary samples i and j having close sentiments, i.e., their sentiment scores can be rounded to the same integer, six representations of sample i (i.e., O_m^i and A_m^i) should be close to the corresponding representations of sample j (i.e., O_n^j and A_n^j). Note that the subscripts m and n represent the modality for sample i and j, respectively. We then form the *inter-sample* positive pairs as

$$\mathcal{P}_2^i = \{(O_m^i, O_n^j), (A_m^i, A_n^j) | m, n \in \{t, v, a\} \ \& \ r(y^i) = r(y^j) \ \& \ i, j \in B \ \& \ i \neq j\},$$

  where y^i denotes the ground-truth of sample i, and $r(\cdot)$ stands for the round function;

- Except for the pairs derived from the above three aspects, the remaining pairs with sample i in the same batch are set as negative pairs $\mathcal{N}_2^i$. Please refer to Appendix A.1 for a more detailed pairing policy.

Specifically, our hierarchical contrastive loss $\mathcal{L}_{con}$ is computed by

$$\mathcal{L}_{con} = \sum_{i \in B} \frac{-1}{|\mathcal{P}_1^i \cup \mathcal{P}_2^i|} \times \sum_{(p,q) \in \mathcal{P}_1^i \cup \mathcal{P}_2^i} \log \frac{f((p,q))}{\sum_{(p',q') \in \text{All Pairs}} f((p',q'))} ,$$

where,

$$f((p,q)) = \exp(\frac{p}{\|p\|_2} \cdot \frac{q}{\|q\|_2} \cdot \frac{1}{\tau}) ,$$

$$\text{All Pairs} = \mathcal{P}_1^i \cup \mathcal{P}_2^i \cup \mathcal{N}_1^i \cup \mathcal{N}_2^i .$$

In the above equation, $|\mathcal{P}_1^i \cup \mathcal{P}_2^i|$ means the number of positive pairs with sample i in a batch B, $(\cdot, \cdot)$ denotes a pair in the corresponding set, e.g., (O_t^i, O_v^i), and τ is a scalar temperature parameter.

The rationale behind this hierarchical contrastive learning subtask is as follows. First, we capture the commonalities across the three modalities within each knowledge type of each sample to reduce the modality gaps under a shared motive. Second, we model the commonalities across samples of close sentiments within each knowledge type to reduce the sample gaps. Third, we capture the differences between the pan-knowledge representations and the knowledge-specific representations in each sample which results in a complementary effect of the two knowledge types of representations. Last but not least, we capture the differences across samples of different sentiments within each knowledge type in order to learn the dynamics of different sentiment intervals.

3.5 Training Procedure

Given the ground truth y and the predictions $\hat{y}$, we can calculate the main MSA task loss by the mean squared error:

$$\mathcal{L}_{task} = \frac{1}{|B|} \sum_{i}^{|B|} (\hat{y}^i - y^i)^2 , \quad (4)$$

where $|B|$ is the number of samples in a batch.

ConKI adopts the learning regime of pretraining followed by fine-tuning. We first pretrain the adapters in ConKI with $\mathcal{L}_{task}$ using an external dataset while fixing the model parameters of the pretrained backbones, considering ConKI only encodes specific knowledge in adapters which have much fewer trainable parameters compared to backbones. Then we fine-tune ConKI with the downstream target dataset by optimizing the overall loss $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda \mathcal{L}_{con} , \quad (5)$$

where λ is a hyperparameter that balances the MSA task loss and the hierarchical contrastive loss. Algorithm 1 shows the full training procedure of ConKI.

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

Table 1: The statistics of CMU-MOSI, CMU-MOSEI and SIMS.

4 Experiments

In this section, we present some experimental details, including datasets, evaluation metrics, baseline models, and experimental results. The implementation details are shown in Appendix B.

4.1 Datasets and Metrics

We conduct experiments on three publicly available benchmark datasets in MSA: CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018) and SIMS (Yu et al., 2020). Table 1 shows the statistics of the datasets. Appendix C describes the details of these datasets.

Following the previous works (Sun et al., 2020; Rahman et al., 2020; Hazarika et al., 2020; Yu et al., 2021; Mai et al., 2021; Han et al., 2021; Yu et al., 2020), we report our experimental results in two forms: regression and classification. For regression, we report mean absolute error (MAE) and Pearson correlation (Corr). For classification, we report binary classification accuracy (Acc-2) and F1 score. Specifically, for CMU-MOSI and CMU-MOSEI datasets, we calculate Acc-2 and F1 scores in negative/positive (zero excluded) and non-negative/positive (zero included) settings as well as seven-class classification accuracy (Acc-7) which shows the percentage of predictions that correctly classified into the same interval of seven intervals between -3 and $+3$. Higher values indicate better performance for all metrics except for MAE.

4.2 Baselines

We compare ConKI with the following state-of-the-art baseline models in MSA: TFN (Zadeh et al.,

Models*	CMU-MOSI					CMU-MOSEI				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
TFN†	0.901	0.698	34.9	-/80.8	-/80.7	0.593	0.700	50.2	-/82.5	-/82.1
LMF†	0.917	0.695	33.2	-/82.5	-/82.4	0.623	0.677	48.0	-/82.0	-/82.1
MulT†	0.861	0.711	-	81.5/84.1	80.6/83.9	0.580	0.703	-	-/82.5	-/82.3
ICCN†	0.862	0.714	39.0	-/83.0	-/83.0	0.565	0.713	51.6	-/84.2	-/84.2
MISA†	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
MAG-BERT†	0.727	0.781	43.62	82.37/84.43	82.50/84.61	0.543	0.755	52.67	82.51/84.82	82.77/84.71
Self-MM†	0.712	0.795	45.79	82.54/84.77	82.68/84.91	0.529	0.767	53.46	82.68/84.96	82.95/84.93
HyCon‡	0.713	0.790	46.6	-/85.2	-/85.1	0.601	0.776	52.8	-/85.4	-/85.6
MMIM†	0.700	0.800	46.65	84.14/86.06	84.00/85.98	**0.526**	0.772	54.24	82.24/85.97	82.66/85.94
ConKI	**0.681**	**0.816**	**48.43**	**84.37/86.13**	**84.33/86.13**	0.529	**0.782**	**54.25**	**82.73/86.25**	**83.08/86.15**

Table 2: Results on CMU-MOSI and CMU-MOSEI. In Acc-2 and F1, the left of the "/" corresponds to "negative/non-negative" and the right corresponds to "negative/positive". *: all models use BERT as the text encoder; †:from (Han et al., 2021); ‡:from (Mai et al., 2021). Best results are marked in bold.

2017), LMF (Liu and Shen, 2018), MulT (Tsai et al., 2019), ICCN (Sun et al., 2020), MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), Self–MM (Yu et al., 2021), HyCon (Mai et al., 2021), and MMIM (Han et al., 2021). The details of these baseline models are shown in Appendix D.

4.3 Results

In accordance with previous work, we run our model five times under the same hyper-parameter settings and report the average performance of all metrics in Table 2 and Table 3. We can observe from these tables that ConKI yields better or competitive results to a range of baseline models on CMU-MOSI, CMU-MOSEI, and SIMS. Specifically, ConKI outperforms all state-of-the-art baseline models in all metrics on CMU-MOSI and SIMS as well as in Corr, Acc-7, Acc-2, F1 scores on CMU-MOSEI. It also achieves closed performance to the best baseline model in MAE on CMU-MOSEI.

Models	MAE	Corr	Acc-2	F1
TFN	0.488	0.496	75.27	75.56
LMF	0.487	0.502	75.36	75.78
MulT	0.485	0.504	75.62	75.84
MISA	0.472	**0.542**	75.49	75.85
MAG-BERT	0.553	0.242	71.43	63.68
Self-MM	0.458	0.535	77.37	77.54
MMIM	0.607	–	69.37	58.00
ConKI	**0.454**	**0.542**	**77.94**	**78.17**

Table 3: Results on SIMS. All baseline model codes are from https://github.com/thuiar/MMSA.

It is notable that the MAE of ConKI on CMU-MOSI outperforms the best baseline model MMIM by around 0.02, which shows ConKI is able to learn effective representations for the MSA task since MAE is the most commonly used evaluation metric in regression tasks. ConKI also presents an excellent performance in the Corr scores on both CMU-MOSI and CMU-MOSEI datasets. The possible reasoning behind this excellent performance is that ConKI uses contrastive learning for recognizing the samples under different sentiments, which could lead to effective ranking results among samples and thus produce a higher Corr score (Swinscow et al., 2002).

Furthermore, Acc-7 of ConKI on CMU-MOSI surpasses the best baseline by 1.78. Though performing classification, especially seven-class classification, is difficult in a regression task, ConKI successfully leverages the contrasts across samples that are classified into seven intervals (by the round function described in Section 3.4) to model the sample dynamics, which brings a great improvement to Acc-7 and Acc-2, demonstrating the efficacy of ConKI in representation learning for MSA. In addition, ConKI shows excellent F1 scores on all datasets, which endorse its potential in real-world applications since F1 is valuable for evaluating imbalanced datasets.

4.4 Ablation Study

We first conduct an ablation study about modalities, as shown in Table 4. We can observe that the inclusion of all three modalities significantly improves the performance of ConKI.

To show the benefits of the proposed knowledge

Models	MAE	Corr	Acc-7	Acc-2	F1
V+A	1.408	0.248	18.72	55.71/54.33	54.37/53.24
T+A	0.700	0.799	48.22	82.45/84.18	82.38/84.16
T+V	0.718	0.798	45.45	82.97/84.88	82.89/84.86
ConKI	**0.681**	**0.816**	**48.43**	**84.37/86.13**	**84.33/86.13**

Table 4: Ablation results when using different modalities.

Models	MAE	Corr	Acc-7	Acc-2	F1
ConKI	**0.681**	**0.816**	**48.43**	**84.37/86.13**	**84.33/86.13**
w/o C1	0.734	0.794	43.56	82.39/84.21	82.35/84.22
w/o C2	0.753	0.789	43.50	82.16/83.84	82.15/83.89
w/o C3	0.710	0.811	44.75	83.29/84.82	83.23/84.80
w/o C4	0.683	0.815	48.05	84.23/86.01	84.16/85.99
w/o $\mathcal{N}_1$	0.689	0.812	47.90	83.88/85.76	83.8/85.74

Table 5: Ablation results of ConKI's components on CMU-MOSI.

injection and hierarchical contrastive learning in ConKI, we conduct a series of ablation experiments on CMU-MOSI, as shown in Table 5 and Table 6. ConKI mainly includes four components: the use of the external dataset (C1), adapters for knowledge injection (C2), pretrained encoders for pan-knowledge (C3), and hierarchical contrastive learning (C4). Table 5 shows that C1 provides advantages by comparing w/o C1 and ConKI. Similarly, C2 provides benefits by comparing w/o C2 and ConKI. C3 is beneficial by comparing w/o C3 and ConKI. C4 is beneficial by comparing w/o C4 and ConKI.

Since the spotlight in our hierarchical contrastive learning is the contrasts between knowledge types, we also compare our model with the model w/o $\mathcal{N}_1$ trained with $\mathcal{L}_{con}$ but without negative pairs $\mathcal{N}_1$, i.e., without disentangling the pan knowledge and specific knowledge. We can conclude that learning differentiated pan-knowledge and knowledge-specific representations is essential in our hierarchical contrastive learning. To better understand the learned pan- and specific-knowledge representations by our hierarchical contrastive learning, we visualize and analyze these representations in Appendix A.2.

To further examine if our performance gain is from the external dataset instead of the proposed knowledge injection and contrastive learning technique, we compare our model with the state-of-the-art baseline models which are fine-tuned by the external dataset. The results from Table 6 show that ConKI still outperforms those baseline models even though they are trained with the external dataset.

Models	MAE	Corr	Acc-7	Acc-2	F1
MISA	0.711	0.804	44.96	82.04/83.99	81.98/84.0
Self-MM	0.712	0.798	45.51	83.59/85.18	83.47/85.12
MMIM	0.716	0.791	45.42	81.81/83.54	81.67/83.46
ConKI	**0.681**†	**0.816**†	**48.43**†	**84.37**†**/86.13**†	**84.33**†**/86.13**†

Table 6: Ablation results when introducing CMU-MOSEI as an external dataset on CMU-MOSI. † means the corresponding result is significantly better than SOTA with p-value < 0.05 based on paired t-test.

Therefore, our gain from ConKI is not solely from adding more data, but from knowledge injection with multi-step transfer learning. Considering the size of CMU-MOSEI is much larger than CMU-MOSI, injecting CMU-MOSEI's knowledge into CMU-MOSI thus has more effects on the downstream task than injecting CMU-MOSI into CMU-MOSEI, as shown in Table 2.

5 Conclusion

In this paper, we present ConKI, a Contrastive Knowledge Injection framework for multimodal sentiment analysis, which learns knowledge-specific representations along with pan-knowledge representations via knowledge injection and hierarchical contrastive learning. ConKI utilizes the pretrained encoders to obtain pan-knowledge representations while generating knowledge-specific representations based on injected adapters that are trained on an external knowledge source. With the specific knowledge, ConKI is able to produce more accurate sentiment predictions than solely using the pan-knowledge representations. To further improve the learning of these representations, we specifically design a hierarchical contrastive learning procedure taking into account the contrasts between knowledge types within each modality, across modalities within one sample, and across samples. Experimental results on three benchmark datasets show that ConKI outperforms all state-of-the-art methods on a range of performance metrics.

Limitations

Our research presents an initial step toward a knowledge injection framework for MSA and still has some limitations to be tackled in the future. Firstly, we can learn more disentangled representations by carefully selecting contrastive pairs for further improvement. Secondly, it will be interest-

ing if we extend our method with multiple external sources that come from different knowledge domains.

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.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 163–171.

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

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Paul Ekman and Erika L. Rosenberg. 2005. What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (facs).

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988.

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.

Anne Lauscher, Olga Majewska, Leonardo FR Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. Association for Computational Linguistics.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2020. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839.

Chen Liu, Yanwei Fu, C. Xu, Siqian Yang, Jilin Li, Chengjie Wang, and Li Zhang. 2021. Learning a few-shot embedding model with contrastive learning. In *AAAI*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhun Liu and Ying Shen. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

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

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 1275–1284, New York, NY, USA. Association for Computing Machinery.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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

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

Thomas Douglas Victor Swinscow, Michael J Campbell, et al. 2002. *Statistics at square one*. Bmj London.

Li Tao, Xueting Wang, and Toshihiko Yamasaki. 2020. Self-supervised video representation learning using inter-intra contrastive framework. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2193–2201.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. Skep: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076.

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.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021b. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. 2021. Knowledge enhanced pretrained language models: A compreshensive survey. *arXiv preprint arXiv:2110.08455*.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal DNN feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33:5812–5823.

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, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, E. Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*.

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

A Hierarchical Contrastive Learning

A.1 Pairing Policy

To further elaborate on the pairing policy of our hierarchical contrastive learning, we show

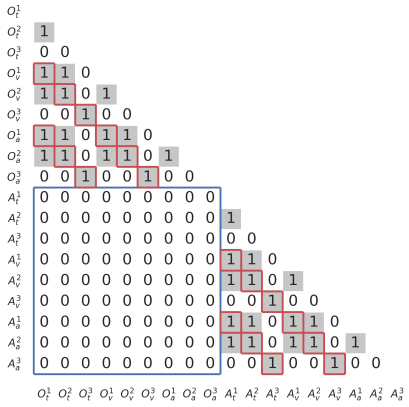

Figure 4: Pairing example of three samples where sample 1 and sample 2 are in the same sentiment interval while sample 3 is in a different sentiment interval. Grey cells with "1" stand for the positive pairs, and white cells with "0" represent the negative pairs.

an example batch that consists of three samples ($\{O_t^i, O_v^i, O_a^i, A_t^i, A_v^i, A_a^i\}_{i=1,2,3} \in B$) where sample 1 and sample 2 belong to the same sentiment interval while sample 3 falls in a different sentiment interval in Figure 4. In this figure, the "1"s in the heatmap represent the positive pairs of the row vectors and column vectors. The "0"s represent the negative pairs of each two vectors.

- From this figure, we can get the *intra-sample* positive pairs as the "1"s with red borders:

$$\begin{aligned}\mathcal{P}_1^B = \{&(O_t^1, O_v^1), (O_t^1, O_a^1), (O_a^1, O_v^1),\\ &(A_t^1, A_v^1), (A_t^1, A_a^1), (A_a^1, A_v^1),\\ &(O_t^2, O_v^2), (O_t^2, O_a^2), (O_a^2, O_v^2),\\ &(A_t^2, A_v^2), (A_t^2, A_a^2), (A_a^2, A_v^2),\\ &(O_t^3, O_v^3), (O_t^3, O_a^3), (O_a^3, O_v^3),\\ &(A_t^3, A_v^3), (A_t^3, A_a^3), (A_a^3, A_v^3)\};\end{aligned}$$

- We represent the *inter-knowledge* negative pairs $\mathcal{N}_1^B$ as the "0"s in the blue zone;

- Since sample 1 and sample 2 have close sentiment scores, we form the *inter-sample* positive pairs as the "1"s without red borders:

$$\begin{aligned}\mathcal{P}_2^B = \{&(O_t^1, O_t^2), (O_t^1, O_v^2), (O_t^1, O_a^2),\\ &(O_v^1, O_t^2), (O_v^1, O_v^2), (O_v^1, O_a^2),\\ &(O_a^1, O_t^2), (O_a^1, O_v^2), (O_a^1, O_a^2),\\ &(A_t^1, A_t^2), (A_t^1, A_v^2), (A_t^1, A_a^2),\\ &(A_v^1, A_t^2), (A_v^1, A_v^2), (A_v^1, A_a^2),\\ &(A_a^1, A_t^2), (A_a^1, A_v^2), (A_a^1, A_a^2)\};\end{aligned}$$

- The remaining white cells with "0" show the negative pairs in $\mathcal{N}_2^B$ which aim to push sample 3 away from sample 1 and sample 2 because they have different sentiments.

A.2 Visualization of Modality Representations

The motivation for us to propose hierarchical contrastive learning into ConKI is that we think modalities will be closer to each other within one sample and across samples in the same sentiment interval and will be far away across samples in different intervals while two knowledge contained in one modality will also be different. We use t-SNE (Van der Maaten and Hinton, 2008) to visualize the distributions of the six representations learned by ConKI with and without hierarchical contrastive learning, as shown in Figure 5.

Though we divide all samples into seven intervals to perform contrastive learning, we take samples of two intervals from the testing set to show the learned representations before and after contrastive learning due to the simplicity of the visualization. From Figure 5 (a), we can easily observe that some of the representations such as the pan-knowledge representations in light blue and the knowledge-specific representations in dark green of samples in two different intervals overlap extremely with each other.

In contrast, these overlapping representations are pushed further in Figure 5 (b) due to sample-level contrasts. It is also obvious that the three knowledge-specific representations of samples in the same interval, e.g., A_t, A_v, A_a of Interval 2 in dark colors and star shape become closer because of both modality-level and sample-level contrasts. Moreover, the distance between the knowledge-specific representations and the pan-knowledge representations, e.g., A_v in dark green and O_v in light green of Interval 2, becomes larger in Figure 5 (b) by knowledge-level contrasts. All of these indicate ConKI is able to perform desired contrastive learning for learning better representations that help improve the performance, even in the generalized scenario, i.e., in the testing set.

B Implementation Details

We use unaligned raw data in all experiments as the previous works (Yu et al., 2021; Han et al., 2021) for fair comparisons. For audio and video modalities, two commonly-used toolkits (COVAREP (Degottex et al., 2014) and Facial Action Coding System (FACS) (Ekman and Rosenberg, 2005)) act as

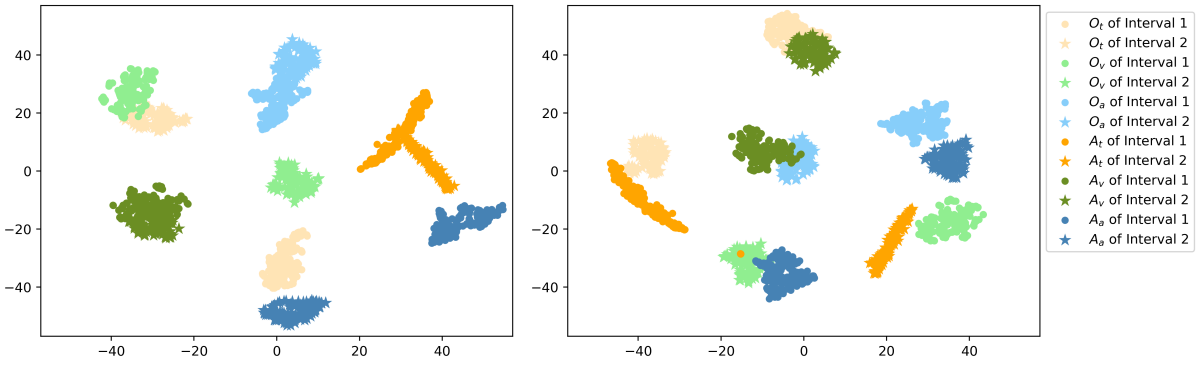

Figure 5: The visualization of the six decomposed representations of samples in the same sentiment interval and different intervals in (a) w/o h-CL; (b) ConKI. In each subfigure, light yellow, light blue, and light green represent pan-knowledge representations of text, audio, and video modalities, respectively, while dark yellow, dark blue, and dark green represent knowledge-specific representations accordingly. Each point or star stands for a sample in Interval 1 or Interval 2.

the feature extractors, respectively. We use uncased 12-layers BERT pretrained model[1] as the text encoder, and two 2-layer transformers as the video and audio encoders, respectively. Adapter$_t$ has three modules inserted before the first, sixth, and eleventh layers of BERT sequentially. Adapter$_v$ and Adapter$_a$ have one module inserted before the second layer of the corresponding encoder. We use CMU-MOSEI as the external dataset for CMU-MOSI and SIMS while using CMU-MOSI for CMU-MOSEI. During the pretraining, the learning rate is set to $5e-5$ and we train for 10 epochs with one epoch for a linear warm-up scheduler. During the fine-tuning of CMU-MOSI and SIMS, the learning rates for encoders and other components are set to $5e-6$ and $1e-6$ respectively with a weight decay 0.001. The temperature parameter τ is set to 0.07 and λ is set to 0.01 after grid-search. We fine tune for 200 epochs with batch size 32. For the fine-tuning of CMU-MOSEI, the learning rates for the text encoder and others are $5e-6$ and $5e-5$, respectively. λ for CMU-MOSEI is set to 0.001. The best performance on the validation dataset is used for testing. We implement our experiments using PyTorch (Paszke et al., 2019) on an Nvidia RTX 2080Ti GPU.

C Datasets

CMU-MOSI is a popular benchmark dataset collected from YouTube. It contains 2,199 video clips sliced from 93 videos where a speaker shares opinions on topics such as movies. Each video is annotated with sentiment scores ranging from -3 (strongly negative) to $+3$ (strongly positive). CMU-MOSEI is the largest MSA dataset that has greater diversity in speakers, topics, and annotations. It contains 22,856 annotated video segments from 1,000 distinct speakers and 250 topics. Each clip also has sentiment scores between $[-3, +3]$. SIMS is a Chinese MSA dataset that contains 2,281 refined video segments. Each sample has one multimodal label and three unimodal labels, with sentiment scores from -1 to $+1$. We translate the Chinese text into English[2] so that we can inject knowledge from English MSA datasets into SIMS. For fair comparisons, all baseline models use the English version to evaluate the performance.

D Baseline Models

TFN. The Tensor Fusion Network (TFN) (Zadeh et al., 2017) encodes three modalities with corresponding embedding subnetworks and uses outer-product to model the unimodal, bimodal, and trimodal interactions as the fusion results.

LMF. The Low-rank Multimodal Fusion (LMF) (Liu and Shen, 2018) utilizes low-rank tensors to improve efficiency of multimodal fusion.

MulT. The Multimodal Transformer (MulT) (Tsai et al., 2019) proposes directional pairwise cross-modal attention that adapts one modality into another for multimodal fusion.

ICCN. The Interaction Canonical Correlation Network (ICCN) (Sun et al., 2020) learns text-based audio and text-based video features by op-

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

[2] https://pypi.org/project/googletrans/

timizing canonical loss. These features are concatenated with the text features for downstream classifiers such as logistic regression.

MISA. The Modality-Invariant and -Specific Representations (MISA) (Hazarika et al., 2020) designs a multitask loss including task prediction loss, reconstruction loss, similarity loss, and difference loss to learn modality-invariant and modality-specific representations.

MAG-BERT. The Multimodal Adaptation Gate for Bert (MAG-BERT) (Rahman et al., 2020) builds an alignment gate that allows audio and video information to leak into the BERT model for multimodal fusion.

Self–MM. The Self-Supervised Multitask Learning (Self–MM) (Yu et al., 2021) proposes a label generation module based on self-supervised learning to obtain unimodal supervision. Then they joint train the multimodal and unimodal tasks for better fusion results.

HyCon. The Hybrid Contrastive Learning (HyCon) (Mai et al., 2021) performs intra- and inter-modal contrastive learning as well as semi-contrastive learning within a modality to explore cross-modal interactions.

MMIM. MultiModal InfoMax (MMIM) (Han et al., 2021) maximizes the mutual information in unimodal input pairs as well as between multimodal fusion result and unimodal input to aid the main MSA task.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
See Limitations section.

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

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

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

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

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

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

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

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
See Table 1 and Appendix C.

C ☑ Did you run computational experiments?

See Section 4.

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

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

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

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
See Section 4.3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
See Appendix B.

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

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.