CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection

Zhen Li, Bing Xu*, Conghui Zhu, Tiejun Zhao
Harbin Institute of Technology, Harbin, China
linklizhen@163.com, {hitxb,conghui,tjzhao}@hit.edu.cn

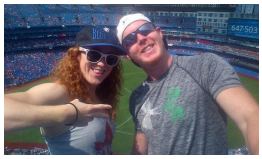
Abstract

Compared with unimodal data, multimodal data can provide more features to help the model analyze the sentiment of data. Previous research works rarely consider token-level feature fusion, and few works explore learning the common features related to sentiment in multimodal data to help the model fuse multimodal features. In this paper, we propose a Contrastive Learning and Multi-Layer Fusion (CLMLF) method for multimodal sentiment detection. Specifically, we first encode text and image to obtain hidden representations, and then use a multi-layer fusion module to align and fuse the token-level features of text and image. In addition to the sentiment analysis task, we also designed two contrastive learning tasks, label based contrastive learning and data based contrastive learning tasks, which will help the model learn common features related to sentiment in multimodal data. Extensive experiments conducted on three publicly available multimodal datasets demonstrate the effectiveness of our approach for multimodal sentiment detection compared with existing methods. The codes are available for use at https://github.com/Link-Li/CLMLF

1 Introduction

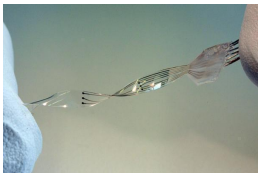
With the development of social networking platforms which have become the main platform for people to share their personal opinions. How to extract and analyze sentiments in social media data efficiently and correctly has broad applications. Therefore, it has attracted attention from both academic and industrial communities (Zhang et al., 2018a; Yue et al., 2019). At the same time, with the increasing use of mobile internet and smartphones, more and more users are willing to post multimodal data (e.g., text, image, and video) about different topics to convey their feelings and sentiments. So multimodal sentiment analysis has become a popular research topic (Kaur and Kautish, 2019).

(a) Heathrow. Fly early tomorrow morning. (positive)



(b) Blue Jays game with the fam! Let's go! (positive)

(c) Ridge Avenue is closed after a partial building collapse and electrical fire Saturday night. (negative)



(d) Flexible spinal cord implants will let paralyzed people walk. (neutral)

Figure 1: Examples of multimodal sentiment tweets

As for multimodal data, the complementarity between text and image can help the model analyze the real sentiment of the multimodal data. As shown in Figure 1, detecting sentiment with only text modality or image modality may not be certain of the true intention of the tweet. Such as Figure 1a, if we only analyze the text modality, we will find that this is a declarative sentence that does not express sentiment. In fact, the girl's smile in the image shows that the sentiment of this tweet is positive. At the same time, in Figure 1c, we can find that the ruins in the image which deepen the expression of negative sentiment in the text.

For multimodal sentiment analysis, we focus on text-image sentiment analysis in social media data. In existing works, some models try to concatenate different modal feature vectors to fuse the multimodal features, such as MultiSentiNet (Xu and Mao, 2017) and HSAN (Xu, 2017). Kumar and Vepa (2020) proposes to use gating mechanism and attention to obtain deep multimodal contextual feature vectors. Multi-view Attentional Network (MVAN) is proposed by Yang et al. (2020) which

* Corresponding author

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2282 - 2294

introduces memory networks to realize the interaction between modalities. Although the above mentioned models are relatively better than unimodal models, the inputs with different modalities are in different vector spaces. Therefore, it is difficult to fuse multimodal data with a simple concatenation strategy, so the improvement is also limited. Furthermore, the gating mechanism and memory network are essentially not designed for multimodal fusion. Although they can help the model analyzes the sentiment in the multimodal data by storing and filtering the features in the data, it is obvious that these methods are difficult to align and fuse the features of text and image. Since Transformers have achieved great success in many fields, such as natural language processing and computer vision (Lin et al., 2021; Khan et al., 2021), we propose **M**ulti-**L**ayer **F**usion (**MLF**) module based on Transformer-Encoder. Benefiting from the multi-headed self-attention in Transformer, which can capture the internal correlation of data vectors. Therefore, text tokens and image patches with explicit and implicit relationships will have higher attention weight allocation to each other which means the MLF module can help align and fuse the token-level text and image features better. And MLF is a multi-layer encoder, which can help improve the abstraction ability of the model and obtain deep features in multimodal data.

Some previous work has explored the application of contrastive learning in the multimodal field. Huang et al. (2021) proposes the application of contrastive learning in multilingual text-to-video search, and Yuan et al. (2021) applies contrastive learning to learn visual representations that embraces multimodal data. However, there is little work to study the application of contrastive learning in multimodal sentiment analysis, so we propose two contrastive learning tasks, **L**abel **B**ased **C**ontrastive **L**earning (**LBCL**) and **D**ata **B**ased **C**ontrastive **L**earning (**DBCL**), which will help the model learn common features related to sentiment in multimodal data. For example, as shown in Figure 1a and Figure 1b. We can find that both tweets show positive sentiment. And we also can find there are smiling expressions in the image of the two tweets which is a common feature of those tweets. If the model can learn common features related to sentiment, it will greatly improve the performance of the model.

In this paper, we propose a **C**ontrastive **L**earning and **M**ulti-**L**ayer **F**usion (**CLMLF**) method for multimodal sentiment analysis based on text and image modalities. For evaluation, CLMLF is verified on three multimodal sentiment datasets, namely MVSA-Single, MVSA-Multiple (Niu et al., 2016) and HFM (Cai et al., 2019). CLMLF achieves better performance compared to several baseline models in all three datasets. Through a comprehensive set of ablation experiments, case study, and visualizations, we demonstrate the advantages of CLMLF for multimodal fusion[1]. Our main contributions are summarized as follows:

- We propose a multi-layer fusion module based on Transformer-Encoder that multi-headed self-attention can help align and fuse token-level features of text and image, and it can also benefit from the depth of MLF which improves model abstraction ability. Experiments show that the proposed architecture of MLF is simple but effective.
- We propose two contrastive learning tasks based on label and data, which leverages sentiment label features and data augmentation. Those two contrastive learning tasks can help the model learn common features related to sentiment in multimodal data, which improve the performance of the model.

2 Approach

2.1 Overview

In this section, we will introduce CLMLF. Figure 2 illustrates the overall architecture of CLMLF model for multimodal sentiment detection that consists of two modules: multi-layer fusion module and multi-task learning module. Specifically, the multi-layer fusion module is on the right in Figure 2, it includes a text-image encoder, image Transformer layer, and text-image Transformer fusion layer modules. The multi-task learning module is on the left in Figure 2, it includes three tasks, sentiment classification, label based contrastive learning and data based contrastive learning tasks.

2.2 Multi-Layer Fusion Module

We use Multi-Layer Fusion module to align and fuse the token-level features of text and image. As shown on the right of Figure 2. First, we

[1]There are also the experimental results and analysis of CLMLF in aspect based multimodal sentiment analysis task, which can refer to Appendix B

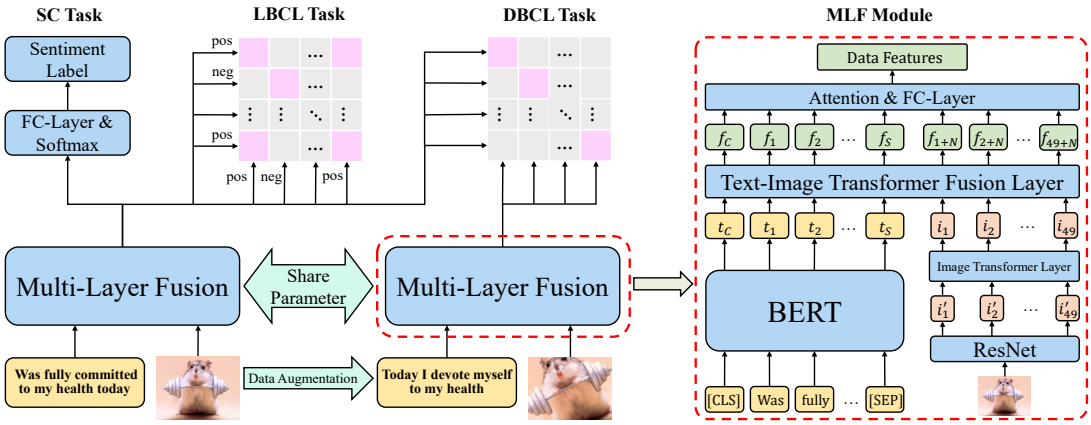


Figure 2: The framework of the proposed CLMLF model

use BERT (Devlin et al., 2019) and ResNet (He et al., 2015) to encode the text and image to obtain the hidden representation of the text $T = \{t_C, t_1, t_2, ..., t_S\}, T \in \mathbb{R}^{n_t \times d_t}$ and the hidden representation of the image $I_c^{'} \in \mathbb{R}^{p_i \times p_i \times d_i}$, and $I_c^{'}$ is the image feature map output by the last layer of convolution layer of ResNet. We transform the hidden representation dimension of $I_c^{'}$ into the same dimension as the T. And we can get the sequence feature representation of the image $I^{'}$ as follows:

$$I^{'} = flatten(I_c^{'} W_I + b_I) \quad (1)$$

Where $I^{'} = \{i_1^{'}, i_2^{'}, ..., i_{n_i}^{'}\}, I^{'} \in \mathbb{R}^{n_i \times d_t}$, $n_i = p_i \times p_i$. And the function of $flatten$ means flatten the input vector by reshaping the first two-dimensions into a one-dimensional.

After that, we will encode the image sequence features $I^{'}$. Here we use the vanilla Transformer-Encoder proposed by Vaswani et al. (2017). Input $I^{'}$ into the image Transformer layer which is based on a multi-layer Transformer-Encoder to obtain the final encoding of image sequence features I.

$$\{i_1, i_2, ..., i_{n_i}\} = TE_I(\{i_1^{'}, i_2^{'}, ..., i_{n_i}^{'}\}) \quad (2)$$
$$I = \{i_1, i_2, ..., i_{n_i}\} \quad (3)$$

Where TE_I means the vanilla Transformer-Encoder of image.

In order to align and fuse the features of text and images, we concatenate the features of the text T and the image sequence features I. We use a new multi-layer Transformer-Encoder as a text-image fusion layer which will align and fuse multimodal features. Then the fusion sequence features of text and image can be obtained. It is as follows:

$$\{f_1, f_2, ..., f_{n_t+n_i}\} = TE_M(concat(T, I)) \quad (4)$$
$$F = \{f_1, f_2, ..., f_{n_t+n_i}\} \quad (5)$$

Where TE_M means the vanilla Transformer-Encoder of multimodal data.

Now, we obtain the sequence features of text and image fusion, but it is obvious that the sequence features can not be used in the classification task. So we use a simple attention layer to get the multimodal representation R.

$$\tilde{q}_i = GELU(f_i W_1 + b_1) W_2 + b_2 \quad (6)$$
$$q_i = exp(\frac{\tilde{q}_i}{\sum_{j=1}^{n_t+n_i} \tilde{q}_j}) \quad (7)$$
$$\tilde{R} = \sum_{i=1}^{n_t+n_i} q_i f_i \quad (8)$$
$$R = GELU(\tilde{R} W_R + b_R) \quad (9)$$

where $GELU$ is the activation function. $R \in \mathbb{R}^{d^t}$

2.3 Sentiment Classification

As shown in the SC task in Figure 2, we feed the above multimodal representation R into the fully connected layer and employ the softmax function for sentiment detection. We use the cross-entropy loss as the classification loss and it is as follows:

$$L_{sc} = Cross\text{-}Entropy(GELU(R W_{sc} + b_{sc})) \quad (10)$$

2.4 Label Based Contrastive Learning

In order to let the model learn the sentiment related features in the multimodal data, we use label based contrastive learning task to help the model extract the sentiment related features while MLF module fuses text and image data. As shown in the LBCL

task in Figure 2, we divide the data in each batch into positive and negative examples according to its sentiment label. For example, in Figure 2, for a negative label of multimodal data, the data in the batch with the same negative labels as positive examples (the square of pink color), and the data with no negative labels are taken as negative examples (the square of gray color).

The specific step can refer to Algorithm 1. The meanings of specific functions in the algorithm are as follows: $einsum$ means Einstein summation convention, $gather$ means gathers values along with an index, and τ represents the contrastive learning's temperature. The algorithm consists of two main steps: the first step is to generate the unmask label L_t according to the data labels in the batch; the second step is to calculate the loss matrix l_{pn}, and use the unmask label L_t and the loss matrix l_{pn} to get the final loss $L_{l\text{-}cl}$, which are the water-red elements in LBCL task on the left in Figure 2.

Algorithm 1 LBCL Algorithm

Require: The sentiment label is L, which is a list of all data in the batch, assuming that the sentiment is divided into three categories: positive (0), neutral (1) and negative (2); The Multi-Layer Fusion Model of MLF; the texts are T; the images are I; C denotes length of L_c; S denotes length of L.

Ensure: Label contrastive learning loss $L_{l\text{-}cl}$

initialize $L_c = [L-0, L-1, L-2]$ and $L_t = list()$
for $i = 1; i <= C; i++$ **do**
  initialize $\tilde{L}_t = list()$
  for $l = 1; l <= T; l++$ **do**
    if $L_c[i][j]$ equals 0 **then**
      $\tilde{L}_t.append(j)$
    end if
  end for
  $L_t.append(\tilde{L}_t)$
end for
$R = MLF(T, I)$
$\tilde{l}_{pn} = einsum(nc, ck- > nk, [R, R^T])$
$l_{pn} = LogSoftmax(l_{pn}/\tau).view(-1)$
$L_{cl} = L_t[L[1]]$
for $q = 2; q <= S, q++$ **do**
  $L_{cl} = concat(L_{cl}, L_t[L[q]] + q \times T)$
end for
$L_{lbcl} = gather(l_{pn}, index = L_{cl})/T$
return L_{lbcl}

2.5 Data Based Contrastive Learning

In order to strengthen the robustness of the model to the data and enhance the learning ability of the model to the invariant features in the data. We add a contrastive learning task based on data augmentation which is DBCL task in Figure 2. Considering the flexible expression of text and images. It may cause the model to be too sensitive to the surface features of data, rather than focus on fusing the invariant features in text and images, that is, effective features. Sentiment related features should exist in these effective features, because the true meaning of the meaning user wants to express should not change with the changes in text and images. For example, both "I had ice cream today. I was very happy" and "I'm very happy today because I ate ice cream" express positive sentiment. The keyword "happy" has not changed which means the happy is an effective feature, but some other words have changed greatly. The data based contrastive learning can force the model learning the effective features in the data, which is more conducive to the model to learn the features related to sentiment in the data. Algorithm 2 describes the process of data based contrastive learning.

Specifically, as for text, we use a data augmentation method called back-translation (Sennrich et al., 2016; Edunov et al., 2018; Xie et al., 2020), which refers to the procedure of translating an existing text x in language E into another language C and then translating it back into E to obtain an augmented text x. As observed by Yu et al. (2018), back-translation can generate diverse paraphrases while preserving the semantics of the original sentences. So we use back-translation to construct positive examples of text in contrastive learning.

For image augmentation, we use a method called RandAugment (Cubuk et al., 2020), which is inspired by AutoAugment (Cubuk et al., 2018). AutoAugment uses a search method to combine all transformations to find a good augmentation strategy. In RandAugment, it does not use search, but instead uniformly samples from the same set of augmentation transformations. In other words, RandAugment is simpler and requires no labeled data as there is no need to search for optimal policies.

2.6 Model Training

The label contrastive loss or data contrastive loss can be simply added to the total loss as a regularization. Can be combined like follows:

Algorithm 2 DBCL Algorithm

Require: The Multi-Layer Fusion Model of MLF; the texts are T; the images are I; BT means back-translation and RA means RandAugment; T denotes of batch size.

Ensure: Data contrastive learning loss $L_{d\text{-}cl}$

$R = MLF(T, I)$
$R_{au} = MLF(BT(T), RA(I))$
$l_{pn} = einsum(nc, ck- > nk, [R, R_{au}^{T}])$
$cl\_label = arange(T)$
$L_{dbcl} = Cross\text{-}Entropy(l_{pn}/\tau, cl\_label)$
return L_{dbcl}

$$L = L_{sc} + \lambda_{lbcl} L_{lbcl} + \lambda_{dbcl} L_{dbcl} \tag{11}$$

where λ_{lbcl} and λ_{dbcl} are coefficients to balance the different training losses.

3 Experimental Setup

3.1 Dataset

We demonstrate the effectiveness of our method on three public datasets which are MVSA-Single, MVSA-Multiple[2] (Niu et al., 2016) and HFM[3] (Cai et al., 2019). Both datasets collect data from Twitter, each text-image pair is labeled by a single sentiment. For a fair comparison, we process the original two MVSA datasets in the same way used in Xu and Mao (2017), as for HFM, we adopt the same data preprocessing method as that of Cai et al. (2019). We randomly split the MVSA datasets into train set, validation set, and test set by using the split ratio 8:1:1. The statistics of these datasets are given in Table 2. The detailed statistics of these datasets are given in Appendix A.

3.2 Implementation Details

For the experiments of CLMLF, we use the Pytorch[4] and HuggingFace Transformers[5] (Wolf et al., 2020) as the implementation of baselines and our method. We use the Bert-base[6] and ResNet-50[7] as the text and image encoder in Multi-Layer Fusion module. The batch size is set to 32, 64 and 128 for MVSA-Single, MVSA-Multiple and HFM. We use AdamW optimizer. The ϵ is 1e-8 and β is (0.9, 0.999). The learning rate is set to 2e-5. Both λ_{lbcl} and λ_{dbcl} are set to 1.0 in Equation 11 during training. For the number of layers of MLF, please refer to Section 4.3. And all the experiments are done on four NVIDIA 3090 GPUs.

3.3 Compared Methods

We compare our model with the unimodal sentiment models and the multimodal baseline models.

Unimodal Baselines: For text modality, CNN (Kim, 2014) and Bi-LSTM (Zhou et al., 2016) are well-known models for text classification tasks. TGNN (Huang et al., 2019) is a text-level graph neural network for text classification. BERT (Devlin et al., 2019) is a pre-trained model for text, and we fine-tuned on the text only. For image modality, OSDA (Yang et al., 2020) is an image sentiment analysis model based on multiple views. ResNet (He et al., 2015) is pre-trained and fine-tuned on the image only.

Multimodal Baselines: MultiSentiNet (Xu and Mao, 2017) is a deep semantic network with attention for multimodal sentiment analysis. HSAN (Xu, 2017) is a hierarchical semantic attentional network based on image captions for multimodal sentiment analysis. Co-MN-Hop6 (Xu et al., 2018) is a co-memory network for iteratively modeling the interactions between multiple modalities. MGNNS (Yang et al., 2021) is a multi-channel graph neural networks with sentiment-awareness for image-text sentiment detection. Schifanella et al. (2016) concatenates different feature vectors of different modalities as multimodal feature representation. Concat(2) means concatenating text features and image features, while Concat(3) has one more image attribute features. MMSD (Cai et al., 2019) fuses text, image, and image attributes with a multimodal hierarchical fusion model. Xu et al. (2020) proposes the D&R Net to fuse text, image, and image attributes by constructing the Decomposition and Relation Network.

4 Results and Analysis

4.1 Overall Result

Table 1 illustrates the performance comparison of our CLMLF model with the baseline methods. We use Weighted-F1 and ACC as the evaluation metrics for MVSA-Single and MVSA-Multiple which is the same as Yang et al. (2021) and use Macro-F1 and ACC as the evaluation metrics for HFM. we

[2]http://mcrlab.net/research/mvsa-sentiment-analysis-on-multi-view-social-data/

[3]https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

[4]https://pytorch.org/

[5]https://github.com/huggingface/transformers

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

[7]https://pytorch.org/vision/stable/models.html

Modality	Model	MVSA-Single		MVSA-Multiple		Model	HFM	
		Acc	F1	Acc	F1		Acc	F1
Text	CNN	0.6819	0.5590	0.6564	0.5766	CNN	0.8003	0.7532
	BiLSTM	0.7012	0.6506	0.6790	0.6790	BiLSTM	0.8190	0.7753
	BERT	0.7111	0.6970	0.6759	0.6624	BERT	0.8389	0.8326
	TGNN	0.7034	0.6594	0.6967	0.6180			
Image	ResNet-50	0.6467	0.6155	0.6188	0.6098	ResNet-50	0.7277	0.7138
	OSDA	0.6675	0.6651	0.6662	0.6623	ResNet-101	0.7248	0.7122
Multimodal	MultiSentiNet	0.6984	0.6984	0.6886	0.6811	Concat(2)	0.8103	0.7799
	HSAN	0.6988	0.6690	0.6796	0.6776	Concat(3)	0.8174	0.7874
	Co-MN-Hop6	0.7051	0.7001	0.6892	0.6883	MMSD	0.8344	0.8018
	MGNNS	0.7377	0.7270	0.7249	0.6934	D&R Net	0.8402	0.8060
	CLMLF	**0.7533**	**0.7346**	0.7200	**0.6983**	CLMLF	**0.8543**	**0.8487**

Table 1: Experimental results of different models on MVSA-Single, MVSA-Multiple and HFM datasets

Dataset	Train	Val	Test	Total
MVSA-S	3611	450	450	4511
MVSA-M	13624	1700	1700	17024
HFM	19816	2410	2409	24635

Table 2: Statistics of the three datasets

have the following observations. First of all, our model is competitive with the other strong baseline models on the three datasets. Second, the multimodal models perform better than the unimodal models on all three datasets. What is more, we found the sentiment analysis on the image modality gets the worst results, this may be that the sentimental features in the image is too sparse and noisy, which makes it difficult for the model to obtain effective features for sentiment analysis. At last, for simple tasks, the performance improvement of multimodal models is limited. For example, on HFM dataset, the improvement of CLMLF relative to BERT is less than MVSA-Single dataset that because HFM is a binary classification task, while MVSA-Single is a three classification task.

We also try to apply CLMLF to aspect based multimodal sentiment analysis task which can refer to Appendix B for details.

4.2 Ablation

We further evaluate the influence of multi-layer fusion module, label based contrastive learning, and data based contrastive learning. The evaluation results are listed in Table 3. The Result shows that the whole CLMLF model achieves the best performance among all models. We can see multi-layer fusion module can improve the performance, which shows that a multi-layer fusion module can fuse the multimodal data. On this foundation, adding the label and data based contrastive learning can improve the model performance more, which means contrastive learning can lead the model to learn common features about sentiment and lead different sentiment data away from each other.

4.3 Influence of MLF Layer

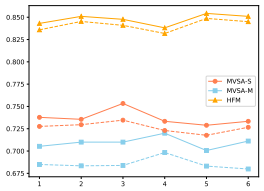
We explored the effects of different layers of Transformer-Encoder on the results. As shown in Figure 3a, fix the image transformer layer and set the text-image transformer fusion layer from 1 to 6. As shown in Figure 3b, fix the text-image transformer fusion layer and set the image transformer layer from 1 to 3. Finally, we selected different combinations of 3-2 (which means three layers of text-image transformer fusion layer and two layers of image transformer layer), 4-2, and 5-1 for the three datasets. This also proves that the contribution of text and images in the dataset is different. It can be seen from Table 1 that CLMLF gains more from the text than images in HFM dataset. Therefore, in MLF module, the layers of transformer related to text are more than images.

4.4 Case Study

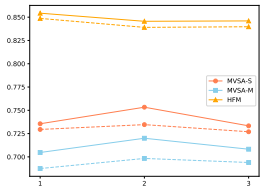
To further demonstrate the effectiveness of our model, we give a case study. We compare the sentiment label predicted based on CLMLF and BERT. As shown in Figure 4, We can find that if we only consider the sentiment of the text, it is difficult to correctly obtain the user's sentimental tendency. For example, for the first data in Figure 4, the meaning of the text is to refer to the image, and

Model	MVSA-Single		MVSA-Multiple		HFM	
	Acc	F1	Acc	F1	Acc	F1
BERT	0.7111	0.6970	0.6759	0.6624	0.8389	0.8326
ResNet-50	0.6467	0.6155	0.6188	0.6098	0.7277	0.7138
+MLF	0.7111	0.7101	0.7059	0.6849	0.8414	0.8355
+MLF, LBCL	0.7378	0.7291	0.7112	0.6863	0.8489	0.8446
+MLF, DBCL	0.7356	0.7276	0.7153	0.6832	0.8468	0.8422
CLMLF	**0.7533**	**0.7346**	**0.7200**	**0.6983**	**0.8543**	**0.8487**

Table 3: Ablation results of our CLMLF



(a) The text-image Transformer fusion layer



(b) The image Transformer layer

Figure 3: Experimental results of different layer of multi-layer fusion module. The solid line indicates the accuracy and the dotted line indicates the F1. The x-axis represents the number of layers of the transformer

the image expresses a positive meaning. for the second data, if we only observe the text, we find that it may express negative sentiments. If add the image, we find that it is just a joke and actually expresses positive sentiment.

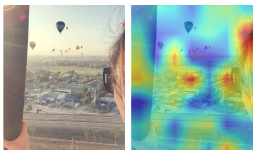
Image	Text	CLMLF	BERT
	Why are you feeling despondent? Take the quiz:	Positive	Neutral
	Thx for taking me to get cheap slushies ?	Positive	Negative
	Car rolls over to avoid real estate sign on Burlington Skyway.	Negative	Neutral

Figure 4: Example of misclassified by BERT and correctly classified by CLMLF

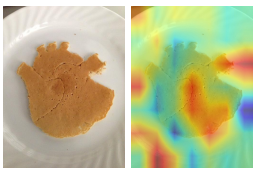
4.5 Visualization

Attention Visualization: We visualize the attention weight of the first head of the Transformer-Encoder in the last layer of the Multi-Layer Fusion module. The result of the attention visualization is shown in Figure 5. We can see that for a given keyword, The model can find the target from the image very well and give it more attention weight. This shows that the model aligns the words in the text with the patch area of the image at a token-level, which plays an important role in the model to fuse text and image features. In particular, for Figure 5b, although "lady" only shows half of the face in the figure, the model still aligns the text and the image very accurately. These indicate that the model aligns the text and image features at token-level according to our assumptions.

(a) The fishing is a little slow but the **flowers** are vibrant and beautiful.



(b) Kimmy, you're one blessed **lady**!



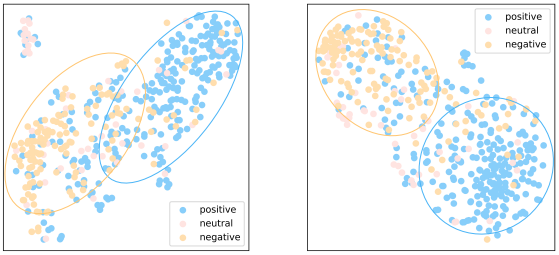
(c) Martha said for Valentine's Day she wanted a heart shaped **pancake** for lunch.

(d) It is truly a **hilarious**, light-hearted read that is a treasure on anyone's bookshelf.

Figure 5: Attention visualization of some multimodal sentiment data examples

Cluster Visualization: In order to verify that our proposed contrastive learning tasks can help the model to learn common features related to sentiment in multimodal data, we conducted a visualization experiment on the MVSA-Single dataset. The data feature vector of the last layer of the model is visualized by dimensionality reduction. We use the TSNE dimensionality reduction algorithm to obtain a 2-dimensional feature vector and visualize it, as shown in Figure 6, Figure 6a is the visualization of the [CLS] of the Bert-base model, and Figure 6b shows the visualization of the fusion result output from the CLMLF model. From the figure, we can

see that after adding contrastive learning, the distance between positive sentiment and negative sentiment in the vector space is greater, and the degree of data aggregation is more obvious. This shows that the model distinguishes these data in vector space according to common features existing in the same sentimental data. Because the number of neutral sentiment data is relatively small, among the visualization results of the two models, CLMLF's visualization results obviously gather the neutral data together, rather than scattered in the vector space like Bert. All these indicate that adding contrast learning can help the model to learn common features related to sentiment which can improve the performance of the model.



(a) BERT (b) Contrastive Learning

Figure 6: Cluster visualization of MVSA-Single

5 Related Work

5.1 Multimodal Sentiment Analysis

In recent years, deep learning models have achieved promising results for multimodal sentiment analysis. MultiSentiNet (Xu and Mao, 2017) and HSAN (Xu, 2017) use LSTM and CNN to encode texts and images to get hidden representations, then concatenate texts and images hidden representations to fuse multimodal features. CoMN (Xu et al., 2018) uses a co-memory network to iteratively model the interactions between visual contents and textual words for multimodal sentiment analysis. Yu et al. (2019) proposes an aspect sensitive attention and fusion network to effectively model the intra-modality interactions including aspect-text and aspect-image alignments, and the inter-modality interactions. MVAN (Yang et al., 2020) applies interactive learning of text and image features through the attention memory network module, and the multimodal feature fusion module is constructed by using a multi-layer perceptron and a stacking-pooling module. Yang et al. (2021) uses multi-channel graph neural networks with sentiment-awareness which is built based on the global characteristics of the dataset for multimodal sentiment analysis.

5.2 Contrastive Learning

Self-supervised learning attracts many researchers for its soaring performance on representation learning in the last several years (Liu et al., 2021; Jing and Tian, 2020; Jaiswal et al., 2021). Many models based on contrastive learning have been proposed in both natural language processing and computer vision fields. ConSERT (Yan et al., 2021), SimCSE (Gao et al., 2021), CLEAR(Wu et al., 2020) proposed the application of contrastive learning in the field of natural language processing. MoCo (He et al., 2020), SimCLR (Chen et al., 2020), SimSiam (Chen and He, 2021), CLIP (Radford et al., 2021) proposed the application of contrastive learning in the field of computer vision, and they also have achieved good results in zero-shot learning and few-shot learning. Recently, contrastive learning has been more and more widely used in the field of multimodality. Huang et al. (2021) uses intra-modal, inter-modal, and cross-lingual contrastive learning which can significantly improves the performance of video search. Yuan et al. (2021) exploits intrinsic data properties within each modality and semantic information from cross-modal correlation simultaneously, hence improving the quality of learned visual representations.

Compared with the above works, we focus on how to align and fuse the token-level features and learn the common features related to sentiment to further improve the performance of model.

6 Conclusion and Future Work

In this paper, we propose a contrastive learning and multi-layer fusion method for multimodal sentiment detection. Compared with previous works, our proposed MLF module performs multimodal feature fusion from the fine-grained token-level, which is more conducive to the fusion of local features of text and image. At the same time, we design learning tasks based on contrastive learning to help the model learn sentiment related features in the multimodal data and improve the ability of the model to extract and fuse features of multimodal data. The experimental results on public datasets demonstrate that our proposed model is competitive with strong baseline models. Especially through visualization, the contrastive learning tasks and multi-layer fusion module we proposed can be ver-

ified with intuitive interpretations. In future work, we will incorporate other modalities such as audio into the sentiment detection task.

Acknowledgements

We are grateful to the anonymous reviewers, meta review and program committee for their insightful comments and suggestions. The work of this paper is funded by the project of National key research and development program of China (No. 2020YFB1406902).

References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

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.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2018. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. corr abs/1512.03385 (2015).

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. 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 3444–3450, Hong Kong, China. Association for Computational Linguistics.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Ramandeep Kaur and Sandeep Kautish. 2019. Multimodal sentiment analysis: A survey and comparison. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET)*, 10(2):38–58.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2021. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751,

Doha, Qatar. Association for Computational Linguistics.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2021. A survey of transformers. *arXiv preprint arXiv:2106.04554*.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999, Melbourne, Australia. Association for Computational Linguistics.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 152–154. IEEE.

Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2399–2402.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 371–378.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jianfei Yu and Jing Jiang. 2019. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414. International Joint Conferences on Artificial Intelligence Organization.

Jianfei Yu, Jing Jiang, and Rui Xia. 2019. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663.

Lei Zhang, Shuai Wang, and Bing Liu. 2018a. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018b. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

A Dataset Statistics

The detailed statistics for the MVSA-Single, MVSA-Multiple and HFM datasets are listed in Table 4. We can see that HFM is a binary classification multimodal sentiment dataset, while MVSA-Single and MVSA-Multiple are three classification multimodal sentiment datasets.

Dataset	Label	Train	Val	Test
MVSA-Single	Positive	2147	268	268
	Neutral	376	47	47
	Negative	1088	135	135
MVSA-Multiple	Positive	9056	1131	1131
	Neutral	3528	440	440
	Negative	1040	129	129
HFM	Positive	8642	959	959
	Negative	11174	1451	1450

Table 4: Number of data for each sentiment category in each dataset

B Aspect Based Multimodal Sentiment

B.1 Experimental Setup

Because CLMLF is designed for sentence-level multimodal sentiment analysis, we have made some minor changes to the input of CLMLF model to adapt to aspect based multimodal sentiment analysis. We change the input form from "[CLS] sentence [SEP]" to "[CLS] sentence [SEP] aspect [SEP]" and no change the input of image modality. Although this change is very simple, CLMLF can work well in aspect based multimodal sentiment analysis tasks and achieves good results.

We use three aspect based multimodal sentiment dataset: Multi-ZOL[8] (Xu et al., 2019), Twitter-15 (Zhang et al., 2018b) and Twitter-17[9] (Lu et al., 2018). The statistics of these datasets are given in Table 5. Compared with the dataset of sentence-level multimodal sentiment analysis, each sentence will have a corresponding aspect attribute. Especially for the Multi-ZOL dataset, each data contains multiple images. And we only randomly select one image for fusion. Although some features are lost, the experimental results show that it is improved compared with the only text modality.

B.2 Results

We compare our model with other baseline models:

[8]https://github.com/xunan0812/MIMN

[9]https://github.com/jefferyYu/TomBERT

Dataset	Train	Val	Test	Total
Multi-ZOL	22743	2843	2843	28429
Twitter-15	3179	1122	1037	5338
Twitter-17	3562	1176	1234	5972

Table 5: Statistics of the three datasets

- LSTM, a standard sentence-level LSTM model without explicitly considering the aspect. Therefore, this result is also the worst.
- AE-LSTM (Wang et al., 2016), an attention-based LSTM for aspect-level sentiment classification, which uses the attention mechanism to capture the important context information related to the aspect.
- RAM (Chen et al., 2017) is a memory based model, which builds memory on the hidden states of a Bi-LSTM and generates aspect representation based on a Bi-LSTM. Then pays multiple attentions on the memory to pick up important information to predict the final sentiment, by combining the features from different attentions non-linearly.
- MIMN (Xu et al., 2019), the multimodal approach for aspect-level sentiment classification task, which adopts multi-hop memory network to model the interactive attention between the aspect word, the textual context, and the visual context.
- TomBERT (Yu and Jiang, 2019), a multimodal model which borrow the idea from self-attention and design a target attention mechanism to perform target-image matching to derive target sensitive visual representations.
- ESAFN (Yu et al., 2019) proposes an entity-sensitive attention and fusion network which capture the intra-modality dynamics by leverages an effective attention mechanism to generate entity-sensitive textual and visual representations. And uses visual attention mechanism to learn the entity-sensitive visual representation. Moreover, ESAFN further fuses the textual and visual representations with a bilinear interaction layer.

Table 6 illustrates the performance comparison of our CLMLF mdoel with the baseline methods. We use Macro-F1 and ACC as the evaluation metrics for all datasets. The experimental results show

Modality	Model	Multi-ZOL		Twitter-15		Twitter-17	
		Acc	F1	Acc	F1	Acc	F1
Text	LSTM	0.5892	0.5729	0.6798	0.5730	0.5592	0.5169
	AE-LSTM	0.5958	0.5895	0.7030	0.6343	0.6167	0.5797
	RAM	0.6018	0.5968	0.7068	0.6305	0.6442	0.6101
	BERT	0.6959	0.6868	0.7387	0.7023	0.6848	0.6553
Multimodal	MIMN	0.6159	0.6051	0.7184	0.6569	0.6588	0.6299
	ESAFN	-	-	0.7338	0.6737	0.6783	0.6422
	TomBERT	-	-	0.7715	0.7175	0.7034	0.6803
	CLMLF	**0.7452**	**0.7075**	**0.7811**	**0.7460**	**0.7098**	**0.6913**

Table 6: Experimental results of different models on aspect based datasets

Model	Multi-ZOL		Twitter-15		Twitter-17	
	Acc	F1	Acc	F1	Acc	F1
BERT	0.6959	0.6868	0.7387	0.7023	0.6848	0.6553
+MLF	0.7301	0.6897	0.7424	0.7017	0.6848	0.6579
+MLF, LBCL	0.7336	0.6953	0.7715	0.7311	0.6969	0.6790
+MLF, DBCL	0.7347	0.7015	0.7445	0.6964	0.6921	0.6722
CLMLF	**0.7452**	**0.7075**	**0.7811**	**0.7460**	**0.7098**	**0.6913**

Table 7: Ablation results of CLMLF

that CLMLF can still achieve good results. We also conducted ablation experiments, as shown in Table 7. The experiments again proved that the multi-layer fusion module, label based contrastive learning task and data based contrastive task we proposed are effective.