Multi-modal Semantic Understanding
with Contrastive Cross-modal Feature Alignment

Ming Zhang[1,2*], **Ke Chang**[1,3*], **Yunfang Wu**[1,3†]

[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Software and Microelectronics, Peking University
[3]School of Computer Science, Peking University
zhangming@stu.pku.edu.cn, {changkegg, wuyf}@pku.edu.cn

Abstract

Multi-modal semantic understanding requires integrating information from different modalities to extract users' real intention behind words. Most previous work applies a dual-encoder structure to separately encode image and text, but fails to learn cross-modal feature alignment, making it hard to achieve cross-modal deep information interaction. This paper proposes a novel CLIP-guided contrastive-learning-based architecture to perform multi-modal feature alignment, which projects the features derived from different modalities into a unified deep space. On multi-modal sarcasm detection (MMSD) and multi-modal sentiment analysis (MMSA) tasks, the experimental results show that our proposed model significantly outperforms several baselines, and our feature alignment strategy brings obvious performance gain over models with different aggregating methods and models even enriched with knowledge. More importantly, our model is simple to implement without using task-specific external knowledge, and thus can easily migrate to other multi-modal tasks. Our source codes are available at https://github.com/ChangKe123/CLFA.

Keywords: contrastive learning, cross-modal feature alignment, sarcasm detection, sentiment analysis

1. Introduction

Recently, with the increasing use of various forms of media such as text, image, video, and audio for information communication on social platforms, semantic understanding involving multi-modal data becomes urgently needed, where sentiment analysis and sarcasm detection are two important tasks.

The multi-modal sentiment analysis task (MMSA) takes data in more than one modal as input, and outputs a classification result with positive, negative and neutral.

The multi-modal sarcasm detection task (MMSD) is more complex and difficult. People often use sarcasm to humorously express mockery, ridicule, criticism, and other emotions towards certain individuals or events. Therefore, sarcasm detection aids in uncovering users' real intentions behind the words, thereby enhancing the effectiveness of semantic understanding. For instance in Figure 1, the sarcasm example contains a twitter "some good stuff here" paired with a crash computer picture. It is likely to misinterpret the intended opinion from the text alone, leading to a biased result.

Current research on multi-modal semantic understanding primarily exploits a dual-encoder structure, which utilizes two separate encoders to process image and text. The extracted features from different modalities are then fused to capture cross-modal information. Previous methods for feature fusion include simple concatenation (Schifanella et al., 2016), attention mechanism (Xu et al., 2020; Pan
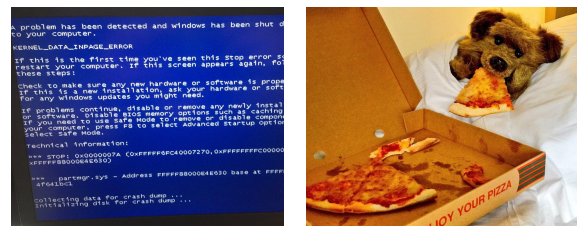
(a) Some good stuff right here.

(b) Fitness whole pizza in my mouth!

Figure 1: Multi-modal sarcastic examples. The left text-image pair conveys sarcasm, while the right has no sarcasm.

et al., 2020) and graph neural networks (Liang et al., 2021; Liu et al., 2022; Liang et al., 2022). However, these methods neglect to consider the semantic alignment between different modalities when fusing multi-modal features. The features derived from image and text data and encoded by separate encoders exist in different semantic spaces. Consequently, utilizing attention or graph neural networks to fuse multi-modal features directly leads to a semantic mismatch, making it hard to learn an effective cross-modal interaction.

In this paper, we propose a novel semantic alignment method for multi-modal deep understanding tasks, namely CLIP-guided **C**ontrastive-**L**earning-based **F**eature **A**lignment (CLFA). Specially, we leverage BERT (Devlin et al., 2019) to encode text and ViT (Dosovitskiy et al., 2021) to encode image. We borrow CLIP (Radford et al., 2021) as

*Equal contribution.

11934

the teacher model and employ contrastive learning to achieve semantic alignment between text and image, projecting the representations derived from different modalities and different encoders into a unified space. We devise a multi-task learning architecture, by treating the feature alignment as an auxiliary task to facilitate the main classification task. In this way, our proposed method can utilize the powerful ability of CLIP and modal-related encoding simultaneously.

We conduct extensive experiments on the public MMSA dataset (Niu et al., 2016) and MMSD dataset (Cai et al., 2019). Our proposed model CLFA brings obvious performance gains over baselines, achieving competitive results with models incorporating external knowledge, including word sentiment, dependency tree and image captions. Besides, we integrate CLFA into a knowledge-enhanced model and several multi-modal fusion methods, achieving better performance and verifying the effectiveness and versatility of our model. More importantly, without using external task-related knowledge and third-party tools, our method can be easily adapted to other multi-modal tasks.

To sum up, our contributions are as follows:

- We propose a novel CLIP-guided contrastive-learning-based architecture for multi-modal semantic alignment, which injects different sorts of features derived from heterogeneous text and image modalities into a unified space.

- Our simple method significantly outperforms baseline models and achieves comparable results with knowledge-enhanced models, on two multi-modal semantic understanding tasks including sentiment analysis and sarcasm detection.

- Plenty of experiments show that our proposed method is effective on different cross-modal aggregating methods. Besides, our method can be combined with other knowledge-based models to get higher performance.

2. Methodology

The overall structure of our proposed model CLFA is illustrated in Figure 2. Specially, the text and image inputs are separately fed into BERT (Devlin et al., 2019) and ViT (Dosovitskiy et al., 2021) to obtain their respective representations. At the core of our model, different representations from text and image are projected into the same deep space guided by the common CLIP representations via contrastive learning, which serves as a subtask to facilitate the main classification task. After that, the text and image features are input into a cross

attention layer to perform multi-modal fusion, and then a classification layer to predict the result.

2.1. Text and Image Encoding

Given an input sentence $S = [s_1, s_2, \cdots, s_n]$, where n represents the length, we adopt BERT to encode each token to the textual representation:

$$f_t(S) = \mathbf{BERT}(S) \tag{1}$$

The last hidden state $f_t(S) = [h_1, h_2, \cdots, h_n] \in \mathcal{R}^{n \times d}$ is considered as the representation of text,

Following previous studies, we employ the pre-trained model ViT for the image modality. The input image is resized to 224×224 and divided into $m = 7 \times 7$ image patches, and then flattened into a sequence $P = [P_1, P_2, \cdots, P_m] \in \mathcal{R}^{m \times 32 \times 32}$. Finally, the image representation $f_i \in \mathcal{R}^{m \times d}$ is obtained through the encoder of ViT:

$$f_i(P) = \mathbf{ViT}(P) \tag{2}$$

We adopt BERT and ViT of the same scale, so the parameter d is unified in our experiment.

2.2. Cross-modal Feature Alignment

As the text and image are encoded by two independent encoders, the distributions of two types of features are inconsistent. As a result, it is difficult to model the interaction between different modalities without alignment. Therefore, we introduce CLIP (Radford et al., 2021) and employ contrastive learning to enhance the model ability to align multi-modal features.

CLIP is pre-trained on a large-scale dataset of image-text pairs, with the aim of feature aligning. In our dual-encoder structure, CLIP innovatively serves as a teacher model to guide the alignment of text and image modalities through contrastive learning.

Our contrastive learning is done in each mini-batch, which contains B text-image pairs, $[(T_1, I_1), (T_2, I_2), \cdots, (T_B, I_B)]$. Using CLIP, we obtain the text representation C_t and image representation C_i, respectively:

$$\begin{aligned} C_t &= [\mathbf{CLIP_t}(T_1), \cdots, \mathbf{CLIP_t}(T_B)] \\ C_i &= [\mathbf{CLIP_i}(I_1), \cdots, \mathbf{CLIP_i}(I_B)] \end{aligned} \tag{3}$$

As CLIP has different output dimension d_C with BERT and ViT, we add a MLP mapping net, to align the output features of BERT and ViT to the same dimension size of CLIP:

$$\begin{aligned} f'_t &= \mathbf{MLP}(f_t) \\ F'_t &= [f'_t(T_1), \cdots, f'_t(T_B)] \\ f'_i &= \mathbf{MLP}(f_i) \\ F'_i &= [f'_i(I_1), \cdots, f'_i(I_B)] \end{aligned} \tag{4}$$
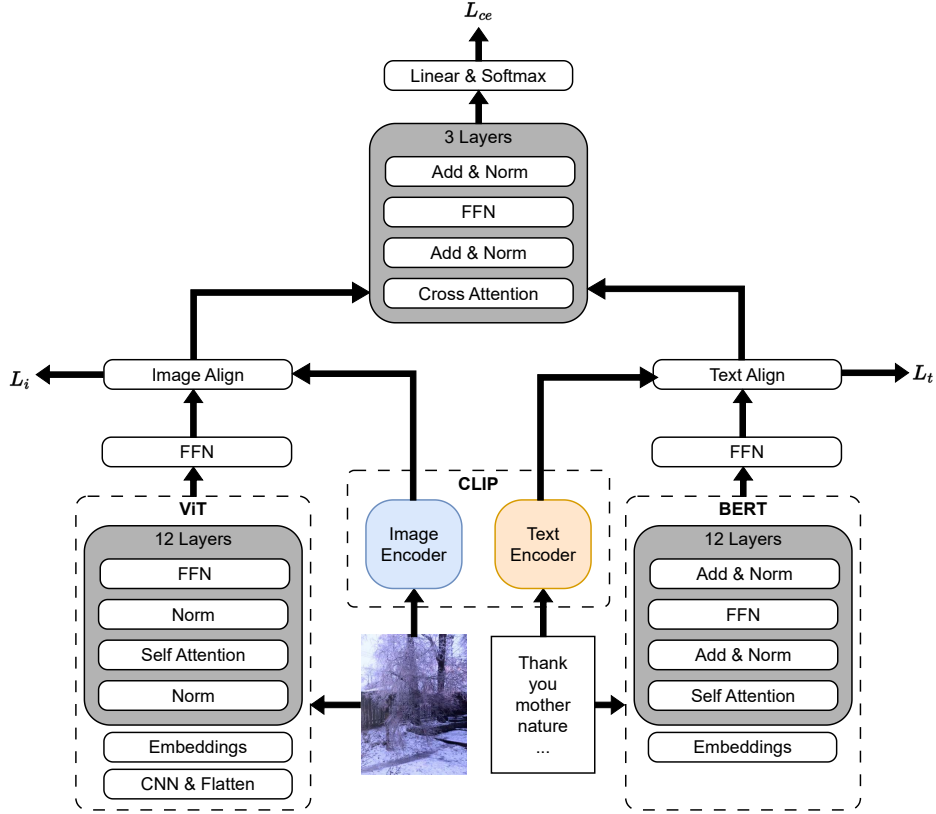
11935

Figure 2: Overview of our proposed model.

That is, the MLP layer takes a d-dim vector as input and outputs a d_C-dim vector.

Image CLIP-ViT representation alignment. For the image alignment, we try to align the CLIP image representation C_i and ViT representation F_i'. Consequently, we consider the matched representations of CLIP and ViT in one mini-batch as positive examples, while unmatched representations as negative examples. Two distinct loss functions with contrastive learning are constructed:

- Using the ViT representation as the anchor, denoted as $\mathcal{L}_{ic}$;

- Using the CLIP representation as the anchor, denoted as $\mathcal{L}_{ci}$.

The loss functions are computed as below:

$$\mathcal{L}_{ic} = -\frac{1}{B} \sum_{k=1}^{B} \log \frac{\mathbf{e}^{\mathbf{sim}(F_{ik}', C_{ik})/\tau}}{\sum_{j=1}^{B} \mathbf{e}^{\mathbf{sim}(F_{ik}', C_{ij})/\tau}}$$
$$\mathcal{L}_{ci} = -\frac{1}{B} \sum_{k=1}^{B} \log \frac{\mathbf{e}^{\mathbf{sim}(C_{ik}, F_{ik}')/\tau}}{\sum_{j=1}^{B} \mathbf{e}^{\mathbf{sim}(C_{ik}, F_{ij}')/\tau}} \quad (5)$$

Here the operation **sim** is to calculate the cosine similarity.

Further, the average of these two loss functions is regarded as the contrastive learning loss function for the image alignment, denoted as $\mathcal{L}_i$:

$$\mathcal{L}_i = \frac{1}{2}\mathcal{L}_{ci} + \frac{1}{2}\mathcal{L}_{ic} \quad (6)$$

Text CLIP-BERT representation alignment. For the text alignment, we also obtain the contrastive learning loss $\mathcal{L}_t$ by aligning the CLIP text representation C_t and BERT representation F_t', in the same way as the image alignment.

Image-Text feature alignment With CLIP as guidance, we perform image-text alignment by combing the loss functions of image alignment and text alignment. The complete contrastive loss function $\mathcal{L}_{con}$ is computed by:

$$\mathcal{L}_{con} = \frac{1}{2}\mathcal{L}_i + \frac{1}{2}\mathcal{L}_t \quad (7)$$

Through this indirect approach, the dual encoders in our model achieve multi-modal feature alignment. It is important to notice that our method is not equivalent to using contrastive learning directly to close the gap between image-text pairs, as in the sarcasm detection task, the twitter text and the image are not always semantically consistent. Actually, the role of the CLIP teacher model is to project features of individual modalities into a unified space and make multi-modal semantics close.

2.3. Cross Attention and Classification

Getting the aligned representations of image and text, we can aggregate them to understand the cross-modal semantics better. For example, we can simply concatenate them and treat the result as a unimodal data for classification. However, in semantic analysis task, the relationship between the two modalities is important and difficult to capture. So we adopt cross attention to discover the complex relationship between text and image, especially finding the inconsistency for the sarcasm detection task.

The main idea in cross attention is using different input to calculate Q and K, V. For multi-modal semantic understanding, we regard the text representation as Query and image representation as Key and Value:

$$Attn(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (8)$$

In the next layer, the F_t' is replaced by the hidden state h of the previous layer:

$$h = Attn(W_q F_t', W_k F_i', W_v F_i') \quad (9)$$

After 3 layers with cross attention, we get the representation and input it into the classification layers, including FFN and softmax.

The final objective function is the weighted sum of the cross-entropy loss and the contrastive learning loss:

$$\mathcal{L} = \alpha \mathcal{L}_{con} + \mathcal{L}_{ce} \quad (10)$$

where α is a hyper-parameter to balance two parts.

3. Experimental Setup

3.1. Dataset and Evaluation Metrics

The publicly available dataset for MMSD was released by Cai et al. (2019), which consists of Twitter image-text pairs. The training data collected twitters containing hashtags such as "`#sarcasm`" as positive samples, and the labels of the validation set and test set were manually verified. The statistical distribution of MMSD dataset is reported in Table 1.

The MVSA dataset (Niu et al., 2016) is a commonly used dataset for the MMSA task. It consists of two parts: Single and Multiple. The Single part is annotated by only one person, and the Multiple part has 3 annotations whose voting results are regarded as golden references by most previous work. The detailed information of MVSA-Single and MVSA-Multipe datasets is listed in Table 2 and Table 3.

The evaluation metrics include accuracy (Acc), precision (P), recall (R), and F1. Following previous researches, we also calculate and report the macro precision, recall and F1. For comparing model performance, we adopt F1 as the main metric.

	Samples	Positive	Negative
Train	19816	8642	11174
Dev	2410	959	1451
Test	2409	959	1450
Total	24635	10560	14075

Table 1: Dataset statistics of sarcasm detection (Cai et al., 2019).

	Samples	Pos.	Neutral	Neg.
Train	3611	2147	376	1088
Dev	450	268	47	135
Test	450	268	47	135
Total	4511	2683	470	1358

Table 2: Dataset statistics of MVSA-Single (Niu et al., 2016).

	Samples	Pos.	Neutral	Neg.
Train	13624	9056	3528	1040
Dev	1700	1131	440	129
Test	1700	1131	440	129
Total	17024	11318	4408	1298

Table 3: Dataset statistics of MVSA-Multiple (Niu et al., 2016).

3.2. Experimental Details

To encode the text, we use bert-base-uncased from HuggingFace. We use Google's ViT-B_32 pre-trained model with the image patch size of 32x32 to obtain the image representation. For feature alignment, we utilize the clip-vit-base-patch32 pre-trained model. The maximum input text length is 77. The Transformer layer with cross-attention is set to 3. The mapping net adopts a two-layer perception network with a hidden dimension of 1536 and an output dimension of 512. During the training, the batch size is set to 8, learning rate is 1e-5, dropout rate is 0.1. The model is trained for 15 epochs. The model employs a warmup strategy with a proportion of 0.1. The model parameter's L2 regularization coefficient is 0.01. The contrastive learning temperature coefficient τ is 0.1.

3.3. Baseline Methods

Since the MMSD task is quite harder, we mainly compare our model with previous methods on this task. As baselines, we do experiments with the image-based models Resnet (He et al., 2016) and

ViT (Dosovitskiy et al., 2021), as well as text-based models BiLSTM and BERT. Additionally, we compare our model with the following multi-modal approaches.

CLIP: (Radford et al., 2021) A multi-modal pre-training model that aligns text and image features through contrastive learning. The aligned text-image features are then concatenated to obtain multi-modal representations. We also experiment a model by combining CLIP and the cross-modal attention, namely **CLIP+Cross**.

MLP+CNN: (Schifanella et al., 2016) It uses MLP to encode text and CNN for image, and the two modalities' features are concatenated and fed into a classification layer for prediction.

HFM: (Cai et al., 2019) A hierarchical fusion model that treats image, text, and image attributes as three modalities. The multi-modal features are summed according to attention weights to obtain the final representation.

D&R Net: (Xu et al., 2020) The Decomposed and Relationship Network decomposes image and text features into HFM same space through a shared layer, and constructs relationships between text and adjective-noun pairs derived from images using attention mechanisms.

ResBert: (Pan et al., 2020) This model utilizes the pretrained ResNet and BERT models. It extracts hashtags from the input text, and fuses text and image modality features using a cross-modal attention mechanism.

InCross: (Liang et al., 2021) By considering the dependency relationships within text and the adjacency relationships within image blocks, this model constructs relationship graphs within and between modalities, and performs multi-modal feature fusion with graph convolutional neural networks.

HKE: (Liu et al., 2022) A knowledge-enhanced hierarchical model that uses image captions as external knowledge. The model captures both atomic-level and compositional-level sentiment incongruity, with attention mechanisms at the atomic level and a graph attention network at the compositional level.

CMGCN: (Liang et al., 2022) This model utilizes an image object recognition tool to identify objects and their attributes. It constructs cross-modal graph relationships based on image object information, and utilizes graph convolutional neural networks and attention mechanisms for feature fusion.

In the recent models, InCross, HKE and CMGCN introduce external knowledge to boost performance. For example, they take the dependency tree as input in the text modality, and leverage generated captions or the object detection result for image. Our model doesn't leverage any external knowledge, and can be combined with these models.

For MMSA task, we compare our method with the baseline model BERT+ViT, which uses BERT to encode text and ViT to encode image, and then concatenates them for classification. Besides, we use RoBERTa pre-trained on Twitter corpus[1] as the text encoder, which is more fit with the MVSA data collected also from the Twitter corpus.

4. Results and Analysis

4.1. Main Results

The overall experimental results are reported in Table 4 and Table 5.

For the MMSD task, our proposed method CLFA achieves an F1 score of 83.91, outperforming all of the previous models without knowledge. Comparing CLFA with the basic model BERT+ViT without feature alignment, we get an F1 improvement with 4.11 points, showing that our alignment strategy can enhance the consistency of different embeddings. Besides, our model is higher than InCross and is competitive with HKE and CMGCN. The three models leverage dependency tree, sentiment of words, and object detection result of images as external knowledge. These comparisons show that our method can improve the understanding ability of the model as the external knowledge can.

Among other models, multi-modal models yield better results than text-based and image-based individual models, suggesting that utilizing only uni-modal information is insufficient for the MMSD task. Different modalities complement each other, and the model needs to learn both modalities and their inter-modal relationships to obtain better semantic understanding. Among uni-modal models, text-based models obviously outperform image-based models, which indicates that the text modality is much more crucial for sarcasm detection as it contains rich semantic information. From the user's perspective, expressing emotions and attitudes in text is more preferred, and the image serves as supplementary information to the text. Therefore, in multi-modal sarcasm detection, the model should focus more on understanding the text modality and utilize the auxiliary information provided by the image modality to improve the detection accuracy. That's also the reason why our model uses the text modality to guide the image modality in the cross attention.

For the MMSA task shown in Table 5 , our method CLFA achieves an F1 score of 72.45 and 68.31 on MVSA-Single and MVSA-Multiple respectively, surpassing the baseline model by 3.61 and 0.92 points, showing that our alignment method is also effective in the multi-modal sentiment analysis task.

---

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base.

Categories	Models	Acc(%)	P(%)	R(%)	F1(%)	Macro		
						P(%)	**R(%)**	**F1 (%)**
Image Based	Resnet*	71.27	63.02	67.36	65.12	70.20	70.61	70.35
	ViT*	72.15	64.72	66.01	65.36	70.97	71.11	71.03
Text Based	BiLSTM*	76.21	71.59	66.74	69.08	75.27	74.61	74.88
	BERT*	79.95	72.2	80.71	76.22	79.18	80.08	79.44
Multi-modal	CLIP*	84.56	**84.57**	74.87	79.42	84.56	82.92	83.53
	CLIP+Cross Attention*	85.14	80.82	82.17	81.49	84.45	84.64	84.54
	MLP+CNN	81.61	-	-	-	79.52	72.47	75.83
	HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	D&R Net	84.02	77.97	83.42	80.60	-	-	-
	ResBert	86.05	78.63	83.31	80.90	78.87	84.46	82.92
	BERT+ViT*	83.73	78.12	81.54	79.80	82.94	83.41	83.15
	Our CLFA*	**86.80**	81.51	**86.44**	**83.91**	**86.09**	**86.74**	**86.36**
With Knowledge	InCross	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	HKE	87.36	81.84	**86.48**	84.09	-	-	-
	CMGCN	**87.55**	**83.63**	84.69	**84.16**	**87.02**	**86.97**	**87.00**

Table 4: Experimental results on MMSD. Those marked with * represent the results of our experiments, while others are retrieved from the published paper. The best results among each group (without knowledge and with knowledge) are highlighted with boldface. The Knowledge category means these model need external knowledge as input, while other models only need image and text themselves.

Datasets	Models	Acc(%)	F1(%)
MVSA-Single	BERT+ViT	69.11	68.84
	Our CLFA	73.11	**72.45**
	RoBERTa+ViT	68.44	68.67
	Our CLFA	**73.33**	72.01
MVSA-Multiple	BERT+ViT	68.14	67.39
	Our CLFA	**69.73**	**68.31**
	RoBERTa+ViT	67.02	65.86
	Our CLFA	69.02	67.26

Table 5: Experimental results on MMSA.

Figure 3: Experiments about α on the dev data of MMSD. Precision, Recall, F1 are macro average values.

On both datasets, CLFA brings obvious performance gains over both the BERT+ViT encoders and RoBERTa+ViT encoders, verifying that our method is effective on different backbone models.

4.2. Hyper-parameter Setting

We conduct experiments on the development set to investigate the coefficient α in Equation 10, as shown in Figure 3. When α is 1, we get the best performance. A larger coefficient of the contrastive learning loss function does not bring further improvement to the model. The alignment of multimodal features can provide assistance for semantic understanding, but excessive pursuit of feature alignment may damage the understanding of text and image by BERT and ViT pre-trained models, leading to a decrease in classification performance.
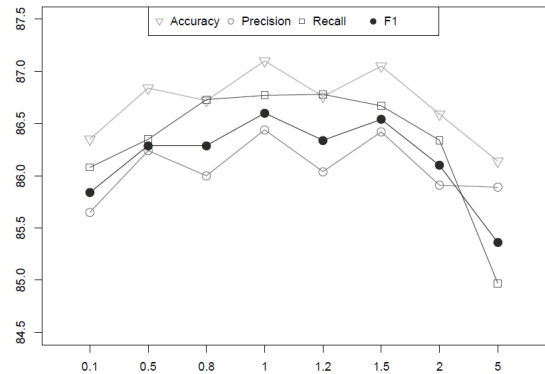
4.3. Visual Analysis on Feature Alignment

For MMSD and MMSA tasks, the input text and image might have different meanings, so we can't directly show the alignment effectiveness on them. But in most cases, the caption of an image has the similar meaning of it. In our experiment, we generate the image captions (textual expressions) using Clipcap (Mokady et al., 2021), and leverage caption-image pairs to show the effectiveness of inter-modal alignment. We don't use the caption in training, so if the model can embedding the caption and image into vectors with high similarity, it can
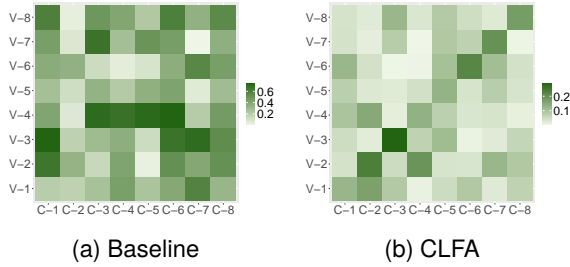
(a) Baseline		(b) CLFA

Figure 4: Text-image alignment heatmap on the MMSD data. C indicates the image caption, V indicates the image. The darker is the color, the higher is the similarity.

perform alignment well.

We present a visual analysis in Figure 4. The sample images are from a mini-batch in the test dataset of MMSD. The figure shows the similarity between text modal features and image modal features, with the paired caption and image on the diagonal. We obtain the similarity value by calculating the cosine distance between the caption and image representations. In the baseline model, there shows no obvious mapping relationship between the image-text pairs. In contrast, in our model, the colors of the diagonal elements are obviously darker than those on the non-diagonal elements, indicating that the image and textual expressions obtain a good alignment with our CLFA method.

4.4. Comparison with Caption Alignment

It should be mentioned that our model CLFA uses CLIP as teacher of encoder, and achieves alignment considering CLIP as a "bridge". To show the necessary of our work, we designed another alignment method:

First, we generate the caption of image by Clip-Cap (Mokady et al., 2021). Second, we directly apply contrastive learning to align the embedding of the caption and the image.

The comparison of CLFA and this method ("Caption" in the table) can be seen in Table 6. CLFA obtains a 2.15 points F1 improvement above Caption alignment, showing that CLIP contains much alignment knowledge and can be a teacher in multi-modal tasks.

Align Methods	Acc(%)	P(%)	R(%)	F1(%)
No Align	83.73	78.12	81.54	79.80
Caption	84.85	79.31	84.36	81.76
CLFA	**86.80**	**81.51**	**86.44**	**83.91**

Table 6: Experimental results with different align methods.

5. Further Experiments with Different Model Settings

5.1. Applying CLFA with Different Aggregating Methods

To investigate whether our CLFA model works on varying model settings, besides cross attention in our main model, we exploit another two multi-modal aggregating methods for further experiments.

Concat It concatenates the two representations derived form text and image modalities.

Co-Attention (Lu et al., 2016) It treats intra-modal and inter-modal attention as different tasks. First, calculating a matrix C as the similarity between text and image modalities:

$$C = \tanh(F_t' W_c F_i') \quad (11)$$

Then, calculating the output with C:

$$h_i = \tanh(W_i F_i' + (W_t F_t')C)$$
$$h_t = \tanh(W_t F_t' + (W_i F_i')C^T) \quad (12)$$
$$h = Concat(h_t, h_i)$$

where W_c, W_t, W_i are trainable parameters.

The Co-Attention method uses C and C^T to measure the relationship between two modalities, and keep the intra-modal information as a linear layer.

Results The experimental results are reported in Table 7. To our expectation, CLFA works well on all of three aggregating methods, obtaining 1.94, 1.07 and 4.11 points F1 improvement on Concat, Co-Attention and Cross-Attention, respectively. The improvement on cross attention is highest, because a well-aligned text-image representation enables cross attention to capture cross-modal relationship. That is, if the input data is well aligned, the cross-attention works better. Besides, the main difficulty in MMSD and MMSA tasks is capturing the inconsistency between modalities, so cross attention with feature alignment achieves better performance.

5.2. Applying CLFA to the Knowledge-enhanced Model

As listed in the Section Baseline Methods, many recent works introduce knowledge to enhance the model's ability for multi-modal semantic understanding. Actually, our model CLFA for feature alignment is orthogonal to the knowledge methods, and can also take external knowledge as input and get higher performance.

To prove this assumption, we design a knowledge-enhanced CLFA, integrating OCR results and word sentiment in the cross-attention layers. Concretely, we modify the three cross-attention layers as below:

Models	Acc(%)	P(%)	R(%)	F1(%)	Macro		
					P(%)	R(%)	F1 (%)
Concat	82.23	75.12	82.79	78.77	81.46	82.33	81.75
Concat+CLFA	84.56	80.29	81.13	80.71	83.86	83.98	83.92
Co-Attention	83.56	77.95	81.86	79.86	82.77	83.27	82.99
Co-Attention+CLFA	84.39	78.76	83.21	80.93	83.62	84.19	83.86
Cross Attention	83.73	78.12	81.54	79.80	82.94	83.41	83.15
Cross Attention+CLFA	**86.80**	**81.51**	**86.44**	**83.91**	**86.09**	**86.74**	**86.36**

Table 7: Experimental results on MMSD with different multi-modal aggregating methods.

Models	Acc(%)	P(%)	R(%)	F1(%)	Macro		
					P(%)	R(%)	F1 (%)
ViT+BERT (Cross Attention)	83.73	78.12	81.54	79.80	82.94	83.41	83.15
+Knowledge	86.38	81.15	85.71	83.37	85.67	86.27	85.92
+Knowledge+CLFA	**87.30**	**81.73**	**87.70**	**84.61**	**86.59**	**87.36**	**86.90**

Table 8: Experimental results on MMSD with knowledge-enhanced methods.

First layer In this layer, we extract sentiment knowledge from SenticNet (Cambria et al., 2022) to better understanding the text data. So we change this layer to do self-attention on the text only:

$$h_1 = Attn'(W_q F'_t, W_k F'_t, W_v F'_t) \quad (13)$$

For every word x in SenticNet, the data contains a value S_x in $[-1, 1]$ to represent the sentiment of it. In our model, we adopt the same way as CMGCN (Liang et al., 2022):

$$SC(x, y) = |S_x - S_y|e^{-S_x S_y}$$
$$Attn'(Q, K, V) = \text{softmax}(\frac{(QK^T) \times (1 + SC)}{\sqrt{d_k}})V \quad (14)$$

Second layer This layer performs text-image cross-attention:

$$h_2 = Attn(W_q h_1, W_k F'_i, W_v F'_i) \quad (15)$$

Since SenticNet can't deal with the image, we use normal attention here.

Third layer We use a third-party OCR tool to extract the expression F'_{OCR} and then compute attention:

$$h_3 = Attn'(W_q h_2, W_k F'_{OCR}, W_v F'_{OCR}) \quad (16)$$

Results The results of knowledge experiments are reported in Table 8. The external knowledge really improves performance by a large margin,

with 3.57 points improvements in F1 score. Even based on the strong knowledge-enhanced model, our method CLFA achieves better results on all metrics, obtaining 1.24 F1 points improvement. It demonstrates the powerful ability of CLFA for multi-modal semantic understanding.

6. Related Work

6.1. Multi-modal Sarcasm Detection

In MMSD task, many neural network models have been proposed. Cai et al. (2019) propose a hierarchical fusion model to integrate different modalities. Xu et al. (2020) propose a Decomposition and Relation Network. that simultaneously learns the relative and relational information between multiple modalities. Pan et al. (2020) emphasize the importance of identifying the inconsistencies within and between modalities. Liang et al. (2021) propose an interactive graph convolutional neural network to learn the emotional inconsistencies within and between modalities. Liu et al. (2022) claim that the inconsistencies of sarcasm exist not only at the atomic level of individual words but also at more complex compositional levels. Liang et al. (2022) use FastRCNN (Girshick, 2015) to identify objects and their attributes in images and then leverage external knowledge to construct an inter-modal graph.

6.2. Multi-modal Sentiment Analysis

In MMSA task, CoMN (Xu et al., 2018) uses memory network to do cross-modal fusion, and Self-MM (Yu et al., 2021) uses joint training on both unimodal task and cross-modal task, which can both be designed with MVSA dataset. Xu and Li

([2023](#)) presents a method using teacher-student framework and extracting features from user comments.

6.3. Multi-modal Alignment

Multi-modal learning aims to overcome the limitations of individual modalities by leveraging the complementary advantages of different modalities, so alignment is an important task in this area. Its advantage lies in the ability to reduce noise and ambiguity from a single modality by combining multiple modalities, enabling the model to have a more comprehensive and accurate understanding of the input data.

The goal of multi-modal alignment is to learn the correspondence between two or more modalities. Currently, there are two main approaches for multi-modal alignment: attention mechanisms and contrastive learning. Yue et al. (2018) employ attention to align different regions in images and text. ViLBERT (Lu et al., 2019) utilizes a Transformer structure with co-attention to align image-text pairs.

On the other hand, contrastive learning has achieved success in coarse-grained multi-modal alignment. Representative works include CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021).

Feature fusion is widely applied for multi-modal learning. The simple method is concatenating or addition of different features to obtain the overall representation. There are also methods based on matrix operations, such as Multi-modal Compact Bilinear (MCB) (Fukui et al., 2016), Multi-modal Low-rank Bilinear (MLB) (Kim et al., 2016), Multi-modal Factorized Bilinear Pooling (MFB) (Yu et al., 2017). What's more, attention-based feature fusion is widely used in VQA tasks, such as co-attention (Yang et al., 2022) and relational attention (Wu et al., 2018).

7. Conclusion

In this paper, we propose a cross-modal feature alignment approach called CLFA. We introduce CLIP as a teacher model in learning multi-modal feature alignment, which enables the model to effectively perform cross-modal interaction during feature fusion. Experiment results show that CLFA gains large improvement on MMSA and MMSD tasks, and can be combined with other aggregating methods and knowledge-enhanced models. Through visual analysis, we can see that CLFA can align text and image well. In conclusion, CLFA is an effective and flexible model which can be used in cross-modal semantic understanding tasks.

Bibliographical 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.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung,

Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4707–4715.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, page 15–27.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.

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.

Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. 2018. Object-difference attention: A simple relational attention for visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 519–527.

Chunpu Xu and Jing Li. 2023. Borrowing human senses: Comment-aware self-training for social media multimodal classification. *arXiv preprint arXiv:2303.15016*.

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, 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.

Zhe Yang, Qingkai Fang, and Yang Feng. 2022. Low-resource neural machine translation with cross-modal alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10134–10146.

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.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Lei Yue, Xin Miao, Pengbo Wang, Baochang Zhang, Xiantong Zhen, and Xianbin Cao. 2018. Attentional alignment networks. In *BMVC*, volume 2, page 7.