1. Introduction

Sarcasm is an interesting and prevailing manner to express users’ opinions [18], which means the real attitude is converse to the literal meaning [19]. With the development of social platforms, sarcasm detection (SD) attracts increasing attention [11, 40, 65] due to its wide application, e.g. product review analysis, political opinion mining [32], etc. Automatically distinguishing sarcastic instances from the mass of non-sarcastic content is important for any online service.

The challenge of multi-modal sarcasm detection (MSD) mainly comes from two aspects. First, the task aims to detect implicit intention from data, which increases the difficulty of learning. Specifically, compared with visual recognition, the expressed attitude of the sarcastic data commonly hides in a normal stimulus and is hard to be identified. Fortunately, the linguistic theory demonstrates that incongruity is an important and effective factor for sarcasm detection [29], which inspires researchers automatically extract the positive and negative seeds [51]. Another challenge lies in that, while both image and text express similar infor-
information is expected in multi-modal tasks [3, 50], this rule is not applicable to SD that discovering dissimilar information. There exists an intrinsic conflict between off-the-shelf techniques for multi-modal learning and the new task in this work.

In order to address the issue, we focus on the inter-modal incongruity for MSD. Sarcasm is a long standing topic in various areas like psychology [43], sociology [54], and neurobiology [31]. Researchers observe that sarcasm occurs when the literal meaning unexpectedly contrasts with the observed facts [22, 43]. The process is defined as counter-factual inference [43]. Besides, the studies from empirical theory [55] find that attitude is another important factor, which is especially effective for obscure cases. In light of these theoretical works, we utilize semantic association and sentiment polarity to verify the incongruity in the sarcasm dataset [6]. As shown in Figure 1, the incongruity of sarcastic data is obviously larger than the non-sarcastic one in factual level, especially in terms of the mean value. Meanwhile, the phenomenon also exists in the affective level. Inspired by the study and verification above, we design our method to detect the incongruity for multi-modal sarcastic data in both the factual and affective levels.

We propose a Dual Incongruity Perceiving (DIP) network, which is consisting of Semantic Intensified Distribution (SID) Modeling and Siamese Sentiment Contrastive (SSC) Learning modules. In SID, based on the semantic association [9, 44], the samples are differentiated by an adaptive strategy. Specifically, we maintain gaussian distributions for sarcastic and non-sarcastic samples respectively, and utilize the probability generated by them to model the incongruity. Since the distributions depend on the extracted embeddings, we introduce a channel-wise reweighting strategy to learn representations related to sarcasm. In SSC, the affective incongruity is perceived by the polarity difference between the image-text pair. To efficiently introduce sentiment information into the network, we employ two siamese layers to transmit knowledge of affective dictionary, i.e. SenticNet. Furthermore, with the help of the polarity intensity, the continuous contrastive learning is proposed to enhance the affective representations. Overall, the factual and affective information are intensified in SID and SSC, and leveraged to explicitly calculate the incongruity for MSD.

Our contributions are three-fold: (1) To our knowledge, DIP is the first work explicitly investigating and modeling incongruity in multi-modal sarcasm detection. (2) It’s a dual perceiving network to learn sarcastic information from factual and affective levels, which utilizes channel-wise reweighting and continuous contrastive strategies to acquire discriminative representations. (3) Extensive comparisons and ablations demonstrate the effectiveness and superiority of the proposed method.

2. Related Work

2.1. Sarcasm Detection

With the rapid development of multimedia, sarcasm becomes prevailing for users to convey the real attitude. In the early stage, researchers detect sarcasm embodied in the text [51, 72]. Roberto et al. utilize hashtags to construct labeled corpus for SD [21]. Riloff et al. develop a bootstrapping method that learns positive and negative phrases respectively [51]. To better utilize the multiple small sarcasm datasets, Guo et al. provide an adversarial model based on latent optimization for transferring knowledge between datasets [23]. Due to the popularity of image-text data, MSD draws increasing attention in recent years [6, 65].

Different from single modal SD, mining the relation between modalities is a crucial tactic for MSD. Schifanella et al. analyze the effectiveness of hand-crafted features and deep representations, then adopting concatenation for the multi-modal prediction [52]. Later, attention-based mechanism becomes the main interaction method of MSD [6, 45]. Cai et al. leverage hierarchical strategy to deeply fuse the representations [6]. Inspired by the significant progress of Transformer, self-attention is employed in MSD to discover the relevance between the modalities [11, 36, 45, 59, 65]. Particularly, realizing the importance of disagreement for MSD, [45, 69] leverage cross-modal attention and expect the modal could implicitly learn the incongruity between images and text. Liu et al. [37] utilize the attention mechanism to model the multi-level i.e.atomic and composition congruity. In order to elaborately take advantage of the mapping between image and text of each instance, graph-based modeling also plays an important role in the recent years [34, 35, 49]. [34] constructs in- and cross-modal graphs to grasp the multi-modal information, and [35] further exploits VQA toolkit [2] to derive the bounding boxes for fine-grained matching.

In the light of DIP, to avoid the complexity of graph-based method [62], we utilize an attention-based strategy for cross-modal interaction. Furthermore, DIP is also different from implicit modeling incongruity [45, 69], where the knowledge exactly learned is unknown [17]. Inspired by the human perception process, we propose a dual perceiving structure to explicitly model the crucial factor incongruity in sarcastic data from factual and affective aspects.

2.2. Sentiment Analysis

Sentiment analysis is closely relevant with SD, which is an attractive topic with widespread application [1, 16, 38]. For visual sentiment analysis, researchers design the handcrafted operators in the early years inspired by psychology and photography [30, 42, 74]. As a typical handcrafted emotional representation, ANP [5] constructs adjective-noun pairs as a descriptor to bridge the mapping between visual
3. Methodology

3.1. Overview

The pipeline of the proposed DIP is illustrated in Figure 2. The image-text pair is formally defined as: $I = \{p^i\}_{i=1}^m$ and $T = \{w^i\}_{i=1}^k$, where $p^i$ represents the $i$-th patch of the image, $w^i$ is the $i$-th word. An image is split into $m$ patches, and a text contains $k$ words. The image and text are first processed by the visual and textual encoders, i.e. ViT [13], BERT [12]. The output embeddings are defined as $e_p \in \mathbb{R}^{m \times C}$ and $e_w \in \mathbb{R}^{k \times C}$, $C$ denotes the number of channels.

To find the informative content, we utilize a cross-modal attention module to implicitly build the interaction between image and text modalities. Specifically, with the embeddings $e_p \in \mathbb{R}^{m \times C}$ and $e_w \in \mathbb{R}^{k \times C}$, we first construct the relation matrix $R \in \mathbb{R}^{m \times k}$ by matrix multiplication in the channel dimension, then pass the matrix through convolution layers:

$$R = \text{Conv}(e_p \cdot e_w^T),$$

where $T$ denotes the transposition. $\text{Conv}$ is implemented by two convolution layers. The large value in $R$ indicates strong relevance. For the visual modality, we add the values of the textual tokens to generate the attention vector $v_p \in \mathbb{R}^m$. The same way is adopted to form the attention vector $v_w \in \mathbb{R}^k$ for the words. Then, $v_p$ and $v_w$ are integrated into $e_p$ and $e_w$ by channel-wise multiplication followed by the sigmoid activation. The aligned visual patch embeddings can be formulated as

$$e_{ap} = \text{sigmoid}(v_p) \cdot e_p,$$

and $e_{aw}$ is processed in the same way. The explicit alignment enforces the image and text representations to be consistent by loss $[3, 50]$. However, sarcasm detection depends on the incongruity between modalities. The strong constraint harms the latent incongruity within the representa-
tions. Therefore, we adopt cross-modal attention to implicitly discover the correspondence regions. Then, the aligned embeddings are sent to SID and SSC.

3.2. Semantic Intensified Distribution Modeling
Counterfactual inference is crucial for perceiving sarcasm [43]. The fact describes the existence of objects or events, which is perceived by semantic information [9, 44]. SID aims to intensify the invariant representations that lead to sarcasm and utilize distributions to model the incongruity in the multi-modal data.

First, we introduce a channel-wise reweighting strategy to learn invariant representations. This strategy is motivated by the observation that some image-text regions are related, but irrelevant to the sarcastic object. Take the example shown in Figure 4 (a), the ‘toast’ exists in both image and text, but actually the key information is ‘egg’. Therefore, inspired by the research about invariant risk minimization [77], we utilize reweighting to find the content most related to sarcasm. Specifically, with the training of the model, representations related to sarcasm are gradually activated by the loss. Furthermore, these embeddings have large variances with different instances [33, 61]. Inspired by this, we propose channel-wise reweighting as:

$$r_p = e_{ap} \cdot \sigma(ReLU(FC(e_{ap}))),$$  \hspace{1cm} (3)

where \(r_p\) denotes reweighted embeddings for patches and \(\sigma\) means the channel-wise variance. The reweighted embeddings for words \(r_w\) is processed in the same way.

After acquiring the discriminative semantic embeddings, we maintain the similarity distributions of sarcastic and non-sarcastic samples, and calculate the probability of multi-modal data belonging to them. Specifically, for the \(r_p\) and \(r_w\), we utilize \(r_v \in \mathbb{R}^C\) and \(r_t \in \mathbb{R}^C\) as [CLS], which are calculated as the average of all the patch and word embeddings. During the training process, we maintain two memory banks \(M_S = \{(r^t_v, r^t_t)\}^q_{i=1}\), \(M_NS = \{(r^t_v, r^t_t)\}^q_{i=1}\) of sarcastic and non-sarcastic semantic representations from previous batches [25], \(q\) represents the length of the memory bank. Based on the observation in Figure 1, we adopt the gaussian distribution, which can be estimated by the following formulas:

$$\mu = \frac{1}{q} \sum_{i=1}^{q} Sim(r^t_v, r^t_t),$$  \hspace{1cm} (4)

$$\sigma = \sqrt{\frac{1}{q} \sum_{i=1}^{q} (Sim(r^t_v, r^t_t) - \mu)^2},$$  \hspace{1cm} (5)

where \(Sim\) denotes the cosine similarity function, \(\mu\) and \(\sigma\) are the mean and variance values of the maintained gaussian distribution. The distributions \(D_s\) and \(D_{ns}\) are denoted as \(D_s \in \mathcal{N}(\mu_s, \sigma_s)\), \(D_{ns} \in \mathcal{N}(\mu_{ns}, \sigma_{ns})\). We model the possibility of the sample belonging to \(D_s\) and \(D_{ns}\) based on the probability density function.

$$p = \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{(Sim(r^t_v, r^t_t) - \mu)^2}{2\sigma^2}},$$  \hspace{1cm} (6)

where \(\tau\) is the temperature controls the importance of \(\sigma\). Next, the factual incongruity \(\lambda_{SID}\) is calculated as \(p_s - p_{ns}\), and utilized to guide MSD. Comparing with adopting the fixed or adaptive threshold to distinguish the sarcastic data, our method leverage gaussian distribution to provide a soft probability. The strategy prevents the bias caused by the hard decision.

3.3. Siamese Sentiment Contrastive Learning
The affective feeling plays a crucial role in MSD [29]. In SSC, we introduce sentiment knowledge to the network, and further model the affective incongruity.

SenticNet [7] is a widely used sentiment dictionary that provides the continuous polarity value of the words. Note that we assign zeros to the words can not be found in SenticNet following [35]. The aligned word embeddings \(e_{aw} \in \mathbb{R}^{k \times C}\) from the cross-attention module are input to the siamese layers to predict the sentiment label for each word. Specifically, the siamese layers consist of a projection head for extracting affective embeddings, and a classifier to obtain polarity value. Then, the text sentiment loss is calculated by the MSE loss,

$$L^{ts} = \frac{1}{k} \sum_{i=1}^{k} (\hat{p}_w^i - p_w^i)^2,$$  \hspace{1cm} (7)

where \(p_w^i\) denotes the polarity value of the \(i\)-th word.

Next, since the embeddings have been implicitly aligned in the cross-modal attention module, we utilize the projection head and classifier with shared parameter to process the images. To further boost the image sentiment representations, we introduce a continuous graph contrastive learning strategy, which constructs continuous supervision labels to capture the intensity of the polarity. In detail, the same as the SID, we use the average of the patch and word embeddings to obtain visual and textual [CLS]. Then \(\hat{p}_v \in \mathbb{R}^B\) and \(\hat{p}_t \in \mathbb{R}^B\) of whole images and texts are obtained by the siamese layers, where \(B\) is the mini-batch size. For an image-text pair, the large polarity difference means the embeddings should be accordingly pushed away. Otherwise, they should be pulled close. Therefore, we construct the supervision \(G_p\) as follows:

$$G_p^i = softmax(exp(-|p_v^i - \hat{p}_v^i|)),$$  \hspace{1cm} (8)

where \(p_v^i, \hat{p}_v^i \in [-1, 1]\) are the intensity of the polarity. The similarity matrix \(G_e\) of embeddings can be calculated by the
Table 1. Comparison with state-of-the-art uni-modal and multi-modal methods on the sarcasm dataset. To fairly and comprehensively verify the performance of the methods, we adopt four backbones i.e. ResNet, ViT, LSTM, and BERT in our experiments. Note HFM* indicates we re-implement HFM on the new backbone.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Method</th>
<th>Acc.</th>
<th>Binary-Average</th>
<th>Macro-Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1</td>
</tr>
<tr>
<td>Image</td>
<td>ResNet [6]</td>
<td>64.76</td>
<td>54.41</td>
<td>70.80</td>
</tr>
<tr>
<td></td>
<td>ViT [13]</td>
<td>67.83</td>
<td>57.93</td>
<td>70.07</td>
</tr>
<tr>
<td></td>
<td>Bi-LSTM [27]</td>
<td>81.90</td>
<td>76.66</td>
<td>70.07</td>
</tr>
<tr>
<td></td>
<td>SIARN [56]</td>
<td>80.57</td>
<td>75.55</td>
<td>75.70</td>
</tr>
<tr>
<td></td>
<td>SMSD [63]</td>
<td>80.90</td>
<td>76.46</td>
<td>75.18</td>
</tr>
<tr>
<td></td>
<td>BERT-Base [12]</td>
<td>83.85</td>
<td>78.72</td>
<td>78.27</td>
</tr>
<tr>
<td></td>
<td>D&amp;R Net [65]</td>
<td>84.02</td>
<td>77.97</td>
<td>80.60</td>
</tr>
<tr>
<td>Text</td>
<td>Bi-LSTM [27]</td>
<td>81.90</td>
<td>76.66</td>
<td>70.07</td>
</tr>
<tr>
<td></td>
<td>SIARN [56]</td>
<td>80.57</td>
<td>75.55</td>
<td>75.70</td>
</tr>
<tr>
<td></td>
<td>SMSD [63]</td>
<td>80.90</td>
<td>76.46</td>
<td>75.18</td>
</tr>
<tr>
<td></td>
<td>BERT-Base [12]</td>
<td>83.85</td>
<td>78.72</td>
<td>78.27</td>
</tr>
<tr>
<td></td>
<td>D&amp;R Net [65]</td>
<td>84.02</td>
<td>77.97</td>
<td>80.60</td>
</tr>
<tr>
<td>Image+Text</td>
<td>HFM [6]</td>
<td>83.44</td>
<td>76.57</td>
<td>84.15</td>
</tr>
<tr>
<td>(ResNet+LSTM)</td>
<td>D&amp;R Net [65]</td>
<td>84.02</td>
<td>77.97</td>
<td>80.60</td>
</tr>
<tr>
<td>Image+Text</td>
<td>HFM* [6]</td>
<td>85.76</td>
<td>82.32</td>
<td>83.88</td>
</tr>
<tr>
<td>(ResNet+BERT)</td>
<td>Res-BERT [45]</td>
<td>84.80</td>
<td>77.80</td>
<td>84.15</td>
</tr>
<tr>
<td></td>
<td>At-BERT [45]</td>
<td>86.05</td>
<td>78.63</td>
<td>83.31</td>
</tr>
<tr>
<td></td>
<td>HKE [37]</td>
<td>87.02</td>
<td>82.97</td>
<td>84.90</td>
</tr>
<tr>
<td></td>
<td>D&amp;R Net [65]</td>
<td>88.20</td>
<td>87.73</td>
<td>82.66</td>
</tr>
<tr>
<td>Image+Text</td>
<td>HFM* [6]</td>
<td>86.63</td>
<td>83.84</td>
<td>84.18</td>
</tr>
<tr>
<td>(ViT+BERT)</td>
<td>InCrossMGSs [34]</td>
<td>86.10</td>
<td>81.38</td>
<td>84.36</td>
</tr>
<tr>
<td></td>
<td>HKE [37]</td>
<td>87.36</td>
<td>81.84</td>
<td>84.48</td>
</tr>
<tr>
<td></td>
<td>CMGCN [35]</td>
<td>87.55</td>
<td>83.63</td>
<td>84.69</td>
</tr>
<tr>
<td></td>
<td>D&amp;R Net [65]</td>
<td>89.59</td>
<td>87.76</td>
<td>86.58</td>
</tr>
</tbody>
</table>

The dot product between the sentiment embeddings \( s_v \in \mathbb{R}^{B \times C} \) and \( s_t \in \mathbb{R}^{B \times C} \), which are the outputs of projection head

\[
G_{e}^{ij} = \text{softmax}(\exp(s_v^i \cdot s_t^j)).
\]  

Besides, the loss of continuous graph contrastive learning is calculated by the Kullback-Leibe (KL) divergence:

\[
\mathcal{L}_{cc} = KL(G_e, G_p). \tag{10}
\]

Then, we use the difference of the sentiment polarity between vision and text as another factor for sarcasm detection, denoted as \( \lambda_{ssc} = |p_v - p_t| \).

After the SID and SSC modules, the embeddings are empirically fused for final prediction. The inter-modal embeddings from the same aspect are processed by element-wise product, then we concatenate the representations from semantic and affective levels. More fusion settings can be found in our ablation experiments.

Considering the incongruity from factual and affective levels, we add the predictions \( y_f \) from fused embedding and the two incongruity factors \( \lambda_{sid} \) and \( \lambda_{ssc} \),

\[
\hat{y} = \text{sigmoid}(y_f + \lambda_{sid} + \lambda_{ssc}). \tag{11}
\]

The binary cross-entropy loss is calculated as:

\[
\mathcal{L}_{bce} = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})]. \tag{12}
\]

Finally, the DIP network for multi-modal sarcasm detection is optimized by the loss:

\[
\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{cc} + \mathcal{L}_{ts}. \tag{13}
\]

4. Experiments

4.1. Dataset and Evaluation Metrics

We conduct experiments on the public MSD dataset [6], and each sample in the dataset consists of an image-text pair. The dataset is divided into training, testing, and validation sets with a ratio of 80%, 10%, and 10%, respectively. During the construction of the dataset, the retrieved tweets with the hashtag #sarcasm are set as positive examples and the others are negative examples. Following previous works [6, 35, 65], we report the accuracy, precision, recall, binary-average, and macro-average results.

4.2. Implementation Details

To ensure fairness, we conduct extensive experiments with various backbones for a comprehensive comparison. In detail, we present the results with ResNet [26], ViT [13], LSTM [27], and BERT [12]. The images are uniformly resized to 224 × 224, and the resolution of a patch is set to 16 in ViT [13]. As a result, the image is split into 196
patches. When utilizing LSTM as the backbone, we adopt Glove [47] for embedding, and the dimension of hidden representations is set as 256. For BERT, we employ the pretrained uncased model. In addition, to unify the dimension of embeddings, we adopt a fully-connected layer following the ResNet and LSTM, which adjusts the output dimension as 768, the same as ViT and BERT. The mini-batch size is set to 16 for experiments with ViT and BERT. Otherwise, the mini-batch is 64. The memory bank stores the latest 256 elements. The network is optimized by stochastic gradient descent with a weight decay of 0.00001. The model is trained for 20 epochs. The learning rate is set to 0.00002 for experiments with ViT and BERT. Otherwise, the learning rate is set to 0.00005 for the rest part. The learning rate is reduced to 0 in the line schedule.

### 4.3. Comparison Methods

We compare DIP with the methods based on image, text, and image+text modalities, respectively.

1) **For the image modality**, we explore the performance of visual information for MSD. Following [35], ResNet [26] and ViT [13] are utilized for comparison.

2) **For the text modality**, we present the performance based on LSTM and BERT. Bi-LSTM [27] is a classical backbone for text analysis. Methods *i.e.* SIARN [56], SMSD [63] designed for single-modal sarcasm detection are also compared. In addition, we fine-tune BERT [12] with the text data and compare with its performance.

3) **For the text+image modality**, we compare all the seven advanced models. HFM [6] designs a hierarchical fusion model to combine information from two modalities. D&R Net [65] uses the semantic association to find the sarcasm clues. Res-BERT [45] and Att-BERT [45] fuse the visual and textual embeddings by concatenation and self-attention mechanism, respectively. InCross-MGs [34] introduces a graph network to depict the image-text pairs. CMGCN [35] builds the connection between regions and words by a cross-modal graph convolutional network (GCN). HKE [37] mines external knowledge to build a hierarchical framework. For a fair comparison, we report the performance of both DIP and the contrastive methods on various backbones.

### 4.4. Comparison with the State-of-the-Art Methods

We conduct extensive experiments to compare DIP with state-of-the-art methods. The results of text, image, and text + image are shown in Table 1. According to the results, we have the following observations.

1) **DIP is clearly superior to single-modal SOTA methods.** Benefit from the complementary information of multi-modal data, DIP improves 21.76% and 5.74% on accuracy compared with visual and textual SOTA methods respectively. On the one hand, compared with only using the data of image modality, it is relatively more effective to detect the sarcasm expressed in the highly semantic text [24]. On the other hand, as an important unit of expressing sarcasm, images can significantly improve the performance of MSD.

2) **Compared with the multi-modal SOTA methods, DIP achieves 2.28%, 1.18%, and 2.04% improvements on accuracy in the three backbone implementation.** For the binary-accuracy precision, DIP improves at least 4.13% compared with the SOTA method. This result demonstrates DIP is particularly adept at recognizing sarcastic data. Moreover, our macro-average metrics also have competitive performance (at least 1.44% improvements), proving that DIP is both effective for distinguishing sarcastic and non-sarcastic data. Furthermore, compared with previous methods implicitly modeling incongruity [37], DIP improves over 2% on accuracy. Therefore, our proposed explicitly modeling factual and affective incongruity method is more effective for MSD.

### 4.5. Ablation Study

To probe the effectiveness of each component in DIP, we conduct ablation experiments. All the experiments are implemented by ViT+BERT. First, we evaluate SID and SSC in Table 2. Base means directly concatenate the [CLS] of visual and textual models. According to the results, we have the following four observations. First, both SID and SSC improve the performance compared with the base model. Second, modeling the incongruity with
Table 3. Comparison of different fusion strategies. M denotes the fusion of image and text embeddings, and L denotes the fusion of semantic and sentiment modules. C, P, S represent concatenation, element-wise product, and element-wise sum respectively. B-F1 and M-F1 denote Binary-F1 and Macro-F1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
<th>B-F1</th>
<th>M-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>88.75</td>
<td>88.41</td>
<td>88.03</td>
</tr>
<tr>
<td>P</td>
<td>89.59</td>
<td>89.04</td>
<td>89.09</td>
</tr>
<tr>
<td>S</td>
<td>89.30</td>
<td>89.00</td>
<td>89.17</td>
</tr>
<tr>
<td>Ours (IA) w/o CR</td>
<td>89.07</td>
<td>85.49</td>
<td>87.68</td>
</tr>
<tr>
<td>Ours (IA) w/o λSID</td>
<td>88.28</td>
<td>86.07</td>
<td>87.98</td>
</tr>
</tbody>
</table>

Table 4. Detailed evaluation of SID. EA means explicit alignment, IA is implicit alignment, CR means channel-wise reweighting.

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
<th>Binary-F1</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (IA)</td>
<td>89.59</td>
<td>87.17</td>
<td>89.01</td>
</tr>
<tr>
<td>Ours (EA)</td>
<td>88.24</td>
<td>85.61</td>
<td>87.84</td>
</tr>
<tr>
<td>Ours (EA) w/o CR</td>
<td>89.91</td>
<td>85.45</td>
<td>87.55</td>
</tr>
<tr>
<td>Ours (EA) w/o λSID</td>
<td>88.03</td>
<td>85.32</td>
<td>87.61</td>
</tr>
<tr>
<td>Ours (IA) w/o CR</td>
<td>89.07</td>
<td>85.49</td>
<td>87.68</td>
</tr>
<tr>
<td>Ours (IA) w/o λSID</td>
<td>88.28</td>
<td>86.07</td>
<td>87.98</td>
</tr>
</tbody>
</table>

$\lambda_{SID}$ achieves higher accuracy, but integrating channel-wise reweighting brings more benefits, which demonstrates the effectiveness of discriminative embeddings. Third, with the help of $L^{cc}$, the binary-F1 is obviously increased. The results demonstrate the effectiveness of continuous contrastive learning for discovering sarcastic data. Fourth, our model combining SID and SSC achieves the best results, showing the components are complementary to each other.

Next, we present the results of different fusion strategies. We empirically evaluate three commonly used fusion methods: concatenation, element-wise sum, and element-wise product. Note M denotes modal (i.e. image, text) fusion, and L represents level (i.e. factual, affective) fusion. As shown in Table 3, the fusion strategy which adopts element-wise product within modalities, and concatenation for the two levels outperforms other strategies. We think this is because the non-linear representation brings more interaction within the modality [75], and the concatenation reserves the information from both semantic and sentiment. Therefore, this form of combination achieves the best performance.

Then, we present the precision of sarcastic and non-sarcastic data by using fused embedding with channel-wise reweighting, incongruity values $\lambda_{SID}$, or $\lambda_{SSC}$. As shown in Figure 3, the fused embedding achieves the best performance. However, the precision of non-sarcastic data is higher than the sarcastic one, which is different from the results of $\lambda_{SID}$ and $\lambda_{SSC}$. Specifically, we find the incongruity values obtained by SID and SSC enable DIP to be more sensitive to the sarcastic data, which improves the performance combined with the fused embedding.

### 4.6. Semantic Effectiveness Analysis

The experimental results of SID are shown in Table 4. We compare the performance of explicit and implicit alignment. The explicit alignment is implemented by a constrastive loss [50] which imposes the similarity of image-text pairs to reach 1. First, DIP with explicit alignment drops 1.35%, 1.56%, 1.17% in accuracy, Binary-F1, and Macro-F1. The phenomenon reflects that explicit alignment impacts the intrinsic incongruity of sarcastic data. Second, the implicit alignment without channel-wise reweighting drops 1.52% in accuracy, which is distinctly larger than the explicit one. The results demonstrate that the process of discovering invariant embedding is relatively more effective for implicit alignment. Different from the explicit strategy which adds a loss term to pull the inter-modal embeddings closer, our method leverages cross attention mechanism to gradually activate the associated patches and words. As a result, the reweighting strategy helps the model to focus on the informative parts. Third, the implicit alignment with $\lambda_{SID}$ is 1.31% higher than without it, but the explicit one only improves 0.21% in accuracy. Based on the observation, we find $\lambda_{SID}$ is more suitable for implicit strategy.

### 4.7. Sentiment Classification Performance

To evaluate the performance of the sentiment module, we conduct experiments to calculate the accuracy of sentiment recognition, as shown in Table 5. The FI [70] and IMDB [41] are commonly used sentiment analysis datasets for image and text respectively. To minimize the effect of the dataset bias [46], we evaluate the methods in two settings. On the one hand, we train DIP on the sarcasm dataset, and test on the sentiment datasets. On the other hand, the predictions of the models pretrained on the sentiment dataset are adopted as labels, which is used to calculate the sentiment recognition accuracy of the DIP. Based on the results, we have the following three observations. First, the base model trained without sentiment branch lacks competitiveness, which needs sentiment supervision information to...
improve the recognition accuracy. Second, we train SID and SSC in a sequential manner, which utilizes the embeddings of SSC for sarcasm detection and sentiment analysis. We find this method is sub-optimal, which may be caused by the MSD is not just relying on the affective cues. Third, our method with continuous contrast learning achieves best performance. Despite the bias among datasets, the experimental results prove that our method learns affective knowledge.

4.8. Visualization

We present some visualization in Figure 4 to further discuss the effectiveness of our method. First, the examples after cross-modal attention are visualized in (a), and the activation maps after channel-wise reweighting are shown in (b). We can observe that the cross-modal attention layers make the network focus on the inter-modal related regions, e.g. eggs and toast, queues, and fence. After the channel-wise reweighting, the model pays more attention to the eggs and fence, which have the closest relation to sarcasm. Second, we provide the samples with high incongruity values for factual and affective respectively. The first two examples in (c) have large factual incongruity values. For instance, the warm and dry are distinctly opposite to the snow in the image. For the next two examples, the sentiment in the text with green color is contrast to the sentiment conveyed in the image. Based on these samples, we can find that both factual and affective incongruity play important roles in MSD.

The examples in (d) show some failure cases of our method. Looking at the left example, both the image and text convey a positive attitude, and there does not exist counterfactual inference. Some cases of sarcasm need strong context knowledge from individuals, which is hard to be differentiated. Observing the right examples, both the low quality of the image and the French may lead to the wrong prediction. Therefore, we think that combining psychology about subjectivity and training a multilingual model may improve the performance of MSD in the future.

5. Conclusion

In this paper, we propose DIP, which learns the incongruity from factual and affective levels. In the factual branch, we design a channel-wise reweighting strategy to focus on the sarcastic regions. Then, the gaussian distribution is utilized to model the incongruity in SID. In the affective branch, we leverage the siamese layers to efficiently introduce sentiment information. Furthermore, the continuous graph contrastive learning is designed to make better use of the intensity of the polarity. Extensive experiments on the MSD dataset indicates that our DIP performs favorably compared with the state-of-the-art methods.

6. Acknowledgments

This work was supported by the National Key Research and Development Program of China Grant (NO. 2018AAA0100400), Natural Science Foundation of Tianjin, China (NO.20JCJQJC00020), and Fundamental Research Funds for the Central Universities.
References


[24] Kaiming He, Xilene Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In CVPR, 2022.


