This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DIP: Dual Incongruity Perceiving Network for Sarcasm Detection

Changsong Wen* Guoli Jia* Jufeng Yang†
TMCC, College of Computer Science, Nankai University, China
downdric@163.com, exped1230@gmail.com, yangjufeng@nankai.edu.cn

Abstract

Sarcasm indicates the literal meaning is contrary to the real attitude. Considering the popularity and complementarity of image-text data, we investigate the task of multi-modal sarcasm detection. Different from other multi-modal tasks, for the sarcastic data, there exists intrinsic incongruity between a pair of image and text as demonstrated in psychological theories. To tackle this issue, we propose a Dual Incongruity Perceiving (DIP) network consisting of two branches to mine the sarcastic information from factual and affective levels. For the factual aspect, we introduce a channel-wise reweighting strategy to obtain semantically discriminative embeddings, and leverage gaussian distribution to model the uncertain correlation caused by the incongruity. The distribution is generated from the latest data stored in the memory bank, which can adaptively model the difference of semantic similarity between sarcastic and non-sarcastic data. For the affective aspect, we utilize siamese layers with shared parameters to learn cross-modal sentiment information. Furthermore, we use the polarity value to construct a relation graph for the mini-batch, which forms the continuous contrastive loss to acquire affective embeddings. Extensive experiments demonstrate that our proposed method performs favorably against state-of-the-art approaches. Our code is released on https://github.com/downdric/MSD.

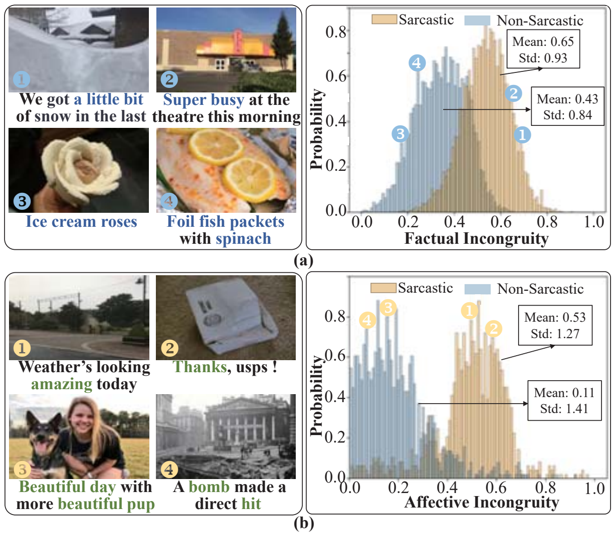

Figure 1. Examples from the sarcasm dataset [6]. (a) shows the samples (left) and statistics (right) for factual incongruity. Acquiring inter-modal semantic similarity S_{inter} from CLIP [50], the factual incongruity is depicted by $1 - S_{inter}$. (b) displays the cases for affective incongruity. Obtaining the models trained on FI (image) [70] and IMDB (text) [41] datasets, the incongruity is represented by the difference in sentiment polarity. For both groups of samples, the top two samples are sarcastic data, and the bottom two samples are non-sarcastic ones.

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-

* Equal contribution.
† Corresponding author.

mation 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 counterfactual 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 facutal 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 tookit [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 hand-crafted 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

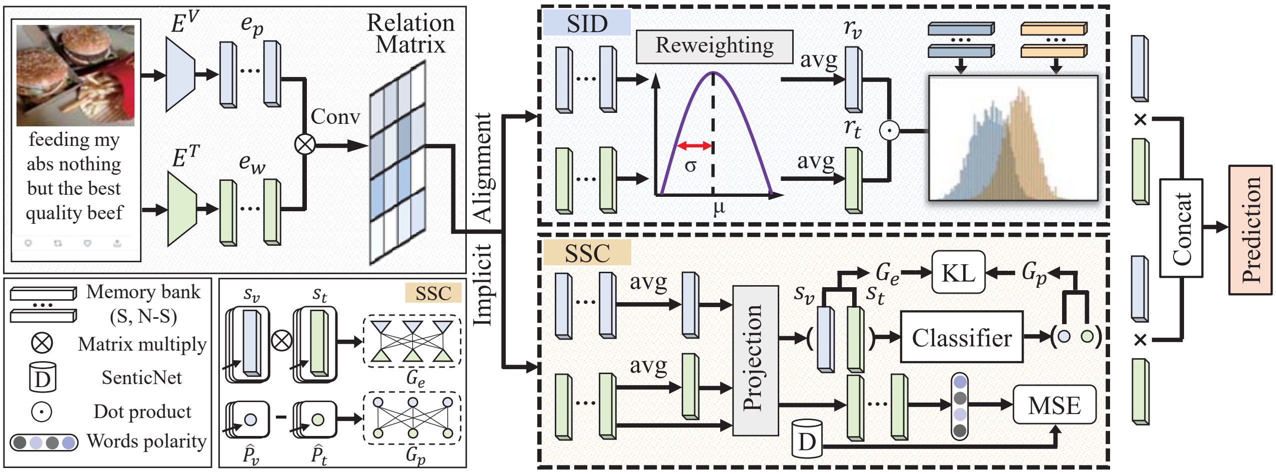

Figure 2. Illustration of the proposed DIP for multi-modal sarcasm detection. The input image and text are processed by ViT and BERT, then delivered to the two branches. The first branch SID leverages reweighting strategy to intensify the regions most related to sarcasm, and utilizes gaussian distribution to discover sarcastic samples. The second branch SSC adopts siamese layers and continuous contrastive learning to jointly learn multi-modal affective embeddings. The process of constructing contrastive graphs G_e and G_p is shown in the sub-graph. Next, embeddings from the dual perception module are fused for sarcasm detection.

concepts and emotion. In recent years, deep neural network (DNN) is leveraged to address the challenge of affective gap [64, 66], ambiguity [28, 60, 67], and emotional region detection [68, 71, 76]. For textual sentiment analysis, researchers automatically construct the sentiment dictionary [53, 57] at the early stage, which can be leveraged to obtain the polarity intensity of words. Later, DNN is utilized to tackle the implicit expression [10, 48], cross-domain [14, 20], and cross-language [4, 8] issues. The DNN methods are powerful on the specific dataset [39], but the dictionaries provide general knowledge without dataset bias [15]. Therefore, introducing word-level domain-invariant features into the DNN is an effective strategy [20]. For multi-modal sentiment analysis, Truong *et al*. leverage images to highlight the salient aspect of the entities [58]. Zhang *et al*. propose a weakly supervised temporal sentiment localization method to detect the parts conveying sentiment [73].

Different from the above works training model on the sentiment datasets, we aim to mine the affective information in the sarcastic samples. Considering the bias between datasets, we design siamese sentiment contrastive learning module to obtain general word-level supervision, which assists the MSD in an end-to-end manner.

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:

$$R = Conv(e_p \cdot {e_w}^T), \tag{1}$$

where T denotes the transposition, $Conv$ 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 formulized as

$$e_{ap} = sigmoid(v_p) \cdot e_p, \tag{2}$$

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 representations. 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:

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

where r_p denotes reweighted embeddings for patches and σ 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 $M_S = \{(r_v^i, r_t^i)\}_{i=1}^q$, $M_{NS} = \{(r_v^i, r_t^i)\}_{i=1}^q$ 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:

$$\mu = \sum_{i=1}^{q} Sim(r_v^i, r_t^i), \tag{4}$$

$$\sigma = \sqrt{\sum_{i=1}^{q} (Sim(r_v^i, r_t^i) - \mu)^2}, \tag{5}$$

where Sim denotes the cosine similarity function, μ and σ 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.

$$p = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\tau(\frac{Sim(r_v^i, r_t^i) - \mu}{\sigma})^2}, \tag{6}$$

where τ is the temperature controls the importance of σ. Next, the factual incongruity λ_{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,

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

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:

$$G_p^{ij} = softmax(exp(-|\hat{p}_v^i - \hat{p}_t^j|)), \tag{8}$$

where $p_v^i, p_t^j \in [-1, 1]$ 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.

Modality	Method	Acc.	Binary-Average			Macro-Average		
			Precision	Recall	F1	Precision	Recall	F1
Image	ResNet [6]	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT [13]	67.83	57.93	70.07	63.43	65.68	71.35	68.40
Text	Bi-LSTM [27]	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN [56]	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD [63]	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT-Base [12]	83.85	78.72	82.27	80.22	81.31	80.87	81.09
Image+Text (ResNet+LSTM)	HFM [6]	83.44	76.57	**84.15**	80.18	79.40	82.45	80.90
	D&R Net [65]	84.02	77.97	83.42	80.60	-	-	-
	DIP	**86.30**	**83.82**	82.35	**83.08**	**85.90**	**85.69**	**85.79**
Image+Text (ResNet+BERT)	HFM* [6]	85.76	82.32	**83.88**	83.09	85.31	85.49	85.27
	Res-BERT [45]	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT [45]	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	HKE [37]	87.02	82.97	84.90	83.92	-	-	-
	DIP	**88.20**	**87.73**	82.66	**85.12**	**88.11**	**87.34**	**87.67**
Image+Text (ViT+BERT)	HFM* [6]	86.63	83.84	84.18	84.01	86.24	86.28	86.26
	InCrossMGs [34]	86.10	81.38	84,36	82.84	85.39	85.80	85.60
	HKE [37]	87.36	81.84	86.48	84.09	-	-	-
	CMGCN [35]	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	DIP	**89.59**	**87.76**	**86.58**	**87.17**	**88.46**	**89.13**	**89.01**

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

$$G_e^{ij} = softmax(exp((s_v^i \cdot s_t^j))). \quad (9)$$

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

$$\mathcal{L}^{cc} = KL(G_e, G_p). \quad (10)$$

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 λ_{SID} and λ_{SSC},

$$\hat{y} = sigmoid(y_f + \lambda_{SID} + \lambda_{SSC}), \quad (11)$$

The binary cross-entropy loss is calculated as:

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

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

$$\mathcal{L} = \mathcal{L}^{bce} + \mathcal{L}^{cc} + \mathcal{L}^{ts}. \quad (13)$$

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

Table 2. Ablation study to prob the SID and SSC utilized in DIP. CR denotes the channel-wise reweighting in SID.

Base	CR	λ_{SID}	$\mathcal{L}^{ts}$	$\mathcal{L}^{cc}$	λ_{SSC}	Acc.	Binary-F1	Macro-F1
✓						85.21	81.42	84.57
✓	✓					87.19	84.86	86.88
✓		✓				86.51	83.14	85.95
✓	✓	✓				88.41	85.70	87.98
✓			✓			86.68	84.04	86.31
✓			✓	✓		87.95	85.21	87.52
✓			✓		✓	87.48	84.18	86.91
✓			✓	✓	✓	88.20	85.70	87.83
✓	✓	✓	✓	✓	✓	**89.59**	**87.17**	**89.01**

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 pre-trained 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 the image and text encoder and 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. InCrossMGs [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

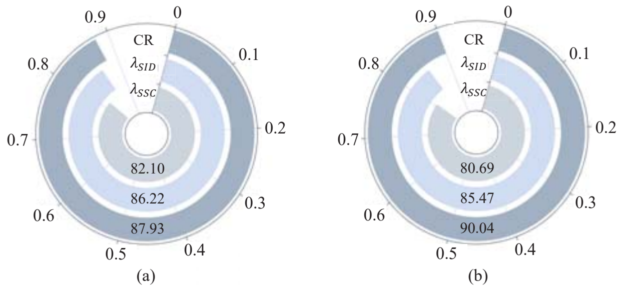

Figure 3. Probing the accuracy of each component of the final predictions. We report the results of precision on both sarcastic and non-sarcastic data in (a) and (b), respectively.

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-average 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.

M \ L	Acc.			B-F1			M-F1		
	C	P	S	C	P	S	C	P	S
C	88.75	88.41	88.03	87.03	85.04	85.55	88.32	87.98	87.67
P	**89.59**	89.04	89.09	**87.17**	86.47	86.85	**89.21**	88.63	88.76
S	89.30	89.00	89.17	86.83	85.79	86.99	88.91	88.41	88.86

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

Method	Acc.	Binary-F1	Macro-F1
Ours (IA)	**89.59**	**87.17**	**89.01**
Ours (EA)	88.24	85.61	87.84
Ours (EA) w/o CR	87.91	85.45	87.55
Ours (EA) w/o λ_{SID}	88.03	85.32	87.61
Ours (IA) w/o CR	88.07	85.49	87.68
Ours (IA) w/o λ_{SID}	88.28	86.07	87.98

λ_{SID} achieves higher accuracy, but integrating channel-wise reweigting brings more benefits, which demonstrates the effectiveness of discriminative embeddings. Third, with the help of $\mathcal{L}^{cc}$, 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 λ_{SID}, or λ_{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 λ_{SID} and λ_{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.

Table 5. Affective recognition accuracy (%) on the FI (image) and IMDB (text) datasets. FI*/IMDB* denotes that training ViT/BERT on the sentiment dataset, then utilizing its prediction on the sarcasm dataset as label for evaluating DIP. MS means the SID and SSC modules are sequentially trained.

Method	Base	w/o $\mathcal{L}^{cc}$	MS	w/o $\mathcal{L}^{cc}$	DIP
FI	55.81	67.10	68.54	69.06	**70.94**
IMDB	52.28	72.55	75.81	77.19	**77.36**
FI*	49.13	65.38	66.29	69.66	**71.03**
IMDB*	52.11	73.00	73.67	76.32	**77.85**

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 constrative 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 λ_{SID} is 1.31% higher than without it, but the explicit one only improves 0.21% in accuracy. Based on the observation, we find λ_{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

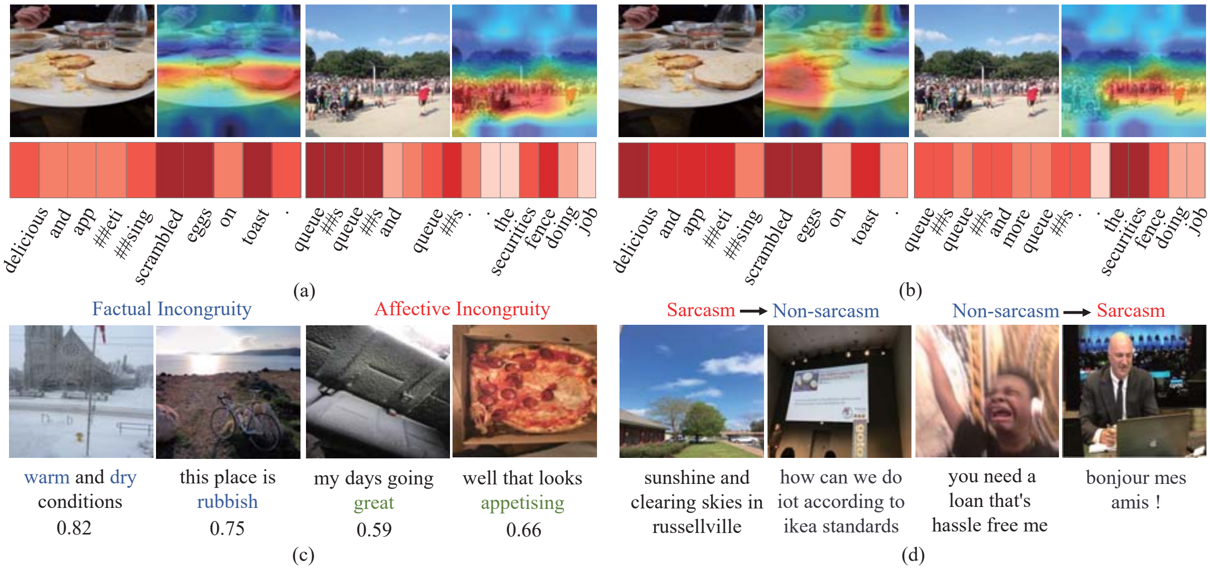

Figure 4. Visualization of our proposed DIP. (a) shows visual and textual attention maps after alignment, (b) shows visual and textual attention maps after invariant reweighting, (c) presents four examples with high incongruity value in semantic and sentiment levels, and (d) displays some failure cases of our method.

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 gaussain 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

[1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *CVPR*, 2021. 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 2

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2, 3

[4] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *LREC*, 2022. 3

[5] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013. 2

[6] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, 2019. 1, 2, 5, 6

[7] Erik Cambria, Yang Li, Frank Z Xing, Soujanya Poria, and Kenneth Kwok. Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *CIKM*, 2020. 4

[8] Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. A sentiment and emotion aware multimodal multiparty humor recognition in multilingual conversational setting. In *COLING*, 2022. 3

[9] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. "factual"or"emotional": Stylized image captioning with adaptive learning and attention. In *ECCV*, 2018. 2, 4

[10] Lingjia Deng and Janyce Wiebe. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *EMNLP*, 2015. 3

[11] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *AAAI*, 2022. 1, 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2018. 3, 5, 6

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5, 6

[14] Abdellah El Mekki, Abdelkader El Mahdaouy, Ismail Berrada, and Ahmed Khoumsi. Domain adaptation for arabic cross-domain and cross-dialect sentiment analysis from contextualized word embedding. In *NAACL*, 2021. 3

[15] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *LREC*, 2006. 3

[16] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: From eye tracking to computational modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1682–1699, 2022. 2

[17] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *CVPR*, 2021. 2

[18] Aniruddha Ghosh and Tony Veale. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *EMNLP*, 2017. 1

[19] Raymond W Gibbs. On the psycholinguistics of sarcasm. *Journal of experimental psychology: General*, 115(1):3, 1986. 1

[20] Chenggong Gong, Jianfei Yu, and Rui Xia. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *EMNLP*, 2020. 3

[21] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: a closer look. In *ACL*, 2011. 2

[22] H Paul Grice. Further notes on logic and conversation. In *Pragmatics*, pages 113–127. Brill, 1978. 2

[23] Xu Guo, Boyang Li, Han Yu, and Chunyan Miao. Latent-optimized adversarial neural transfer for sarcasm detection. In *NAACL*, 2021. 2

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 6

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 4

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5, 6

[28] Guoli Jia and Jufeng Yang. S2-ver: Semi-supervised visual emotion recognition. In *ECCV*, 2022. 3

[29] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing context incongruity for sarcasm detection. In *ACL*, 2015. 1, 4

[30] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011. 2

[31] Christopher M Kipps, Peter J Nestor, Julio Acosta-Cabronero, Robert Arnold, and John R Hodges. Understanding social dysfunction in the behavioural variant of frontotemporal dementia: the role of emotion and sarcasm processing. *Brain*, 132(3):592–603, 2009. 2

[32] Y Alex Kolchinski and Christopher Potts. Representing social media users for sarcasm detection. In *EMNLP*, 2018. 1

[33] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. In *ICLR*, 2022. 4

[34] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *ACM MM*, 2021. 2, 5, 6

[35] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *ACL*, 2022. 2, 4, 5, 6

[36] Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *EMNLP*, 2022. 2

[37] Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *EMNLP*, 2022. 2, 5, 6

[38] Shengzhe Liu, Xin Zhang, and Jufeng Yang. SER30K: A large-scale dataset for sticker emotion recognition. In *ACM MM*, 2022. 2

[39] Xin Liu, Guobao Xiao, Riqing Chen, and Jiayi Ma. Pgfnet: Preference-guided filtering network for two-view correspondence learning. *IEEE Transactions on Image Processing*, 32:1367–1378, 2023. 3

[40] Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *EMNLP*, 2021. 1

[41] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011. 1, 7

[42] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 2010. 2

[43] Skye McDonald. Exploring the process of inference generation in sarcasm: A review of normal and clinical studies. *Brain and language*, 68(3):486–506, 1999. 2, 4

[44] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 2021. 2, 4

[45] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *EMNLP*, 2020. 2, 5, 6

[46] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *ECCV*, 2018. 7

[47] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 6

[48] Minh Hieu Phan and Philip O Ogunbona. Modelling context and syntactical features for aspect-based sentiment analysis. In *ACL*, 2020. 3

[49] Joan Plepi and Lucie Flek. Perceived and intended sarcasm detection with graph attention networks. In *EMNLPW*, 2021. 2

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 7

[51] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, 2013. 1, 2

[52] Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *ACM MM*, 2016. 2

[53] Fabrizio Sebastiani and Andrea Esuli. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, 2006. 3

[54] Amy Sparks, Skye McDonald, Bianca Lino, Maryanne O'Donnell, and Melissa J Green. Social cognition, empathy and functional outcome in schizophrenia. *Schizophrenia research*, 122(1-3):172–178, 2010. 2

[55] Dan Sperber and Deirdre Wilson. Précis of relevance: Communication and cognition. *Behavioral and brain sciences*, 10(4):697–710, 1987. 2

[56] Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. Reasoning with sarcasm by reading in-between. In *ACL*, 2018. 5, 6

[57] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010. 3

[58] Quoc-Tuan Truong and Hady W Lauw. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *AAAI*, 2019. 3

[59] Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. Multimodal sarcasm target identification in tweets. In *ACL*, 2022. 2

[60] Lijuan Wang, Guoli Jia, Ning Jiang, Haiying Wu, and Jufeng Yang. Ease: Robust facial expression recognition via emotion ambiguity-sensitive cooperative networks. In *ACM MM*, 2022. 3

[61] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. Implicit semantic data augmentation for deep networks. In *NeurIPS*, 2019. 4

[62] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020. 2

[63] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *WWW*, 2019. 5, 6

[64] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *CVPR*, 2022. 3

[65] Nan Xu, Zhixiong Zeng, and Wenji Mao. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *ACL*, 2020. 1, 2, 5, 6

[66] Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. Solver: Scene-object interrelated visual emotion reasoning network. *IEEE Transactions on Image Processing*, 30:8686–8701, 2021. 3

[67] Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. A circular-structured representation for visual emotion distribution learning. In *CVPR*, 2021. 3

[68] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 2018. 3

[69] Zhangmingjia Yin and Fucheng You. Multimodal sarcasm semantic detection based on inter-modality incongruity. In *ICCAID*, 2022. 2

[70] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 2016. 1, 7

[71] Haimin Zhang and Min Xu. Weakly supervised emotion intensity prediction for recognition of emotions in images. *IEEE Transactions on Multimedia*, 23:2033–2044, 2020. 3

[72] Meishan Zhang, Yue Zhang, and Guohong Fu. Tweet sarcasm detection using deep neural network. In *COLING*, 2016. 2

[73] Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, 2022. 3

[74] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 2014. 2

[75] Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73, 2021. 7

[76] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6729–6751, 2021. 3

[77] Xiao Zhou, Yong Lin, Renjie Pi, Weizhong Zhang, Renzhe Xu, Peng Cui, and Tong Zhang. Model agnostic sample reweighting for out-of-distribution learning. In *ICML*, 2022. 4