Mutual-Enhanced Incongruity Learning Network for Multi-Modal Sarcasm Detection

Yang Qiao[1] ,Liqiang Jing[1], Xuemeng Song[1*], Xiaolin Chen[2], Lei Zhu[3], Liqiang Nie[4]

[1]School of Computer Science and Technology, Shandong University, Qingdao, China
[2]School of Software, Shandong University, Jinan, China
[3]School of Information Science and Engineering, Shandong Normal University, Jinan, China
[4]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen campus), Shenzhen, China
yang.qiao@mail.sdu.edu.cn, {jingliqiang6,sxmustc,cxlicd,leizhu0609,nieliqiang}@gmial.com

Abstract

Sarcasm is a sophisticated linguistic phenomenon that is prevalent on today's social media platforms. Multi-modal sarcasm detection aims to identify whether a given sample with multi-modal information (*i.e.*, text and image) is sarcastic. This task's key lies in capturing both inter- and intra-modal incongruities within the same context. Although existing methods have achieved compelling success, they are disturbed by irrelevant information extracted from the whole image and text, or overlooking some important information due to the incomplete input. To address these limitations, we propose a Mutual-enhanced Incongruity Learning Network for multi-modal sarcasm detection, named MILNet. In particular, we design a local semantic-guided incongruity learning module and a global incongruity learning module. Moreover, we introduce a mutual enhancement module to take advantage of the underlying consistency between the two modules to boost the performance. Extensive experiments on a widely-used dataset demonstrate the superiority of our model over cutting-edge methods.

Introduction

Sarcasm is a special linguistic phenomenon that aims to express people's emotions contrary to the normal interpretations, which frequently appears on online social media platforms. Therefore, sarcasm detection is particularly important in customer service, opinion mining, and various tasks that require understanding people's emotions, and has gained increasing research attention. Early sarcasm detection methods (Davidov, Tsur, and Rappoport 2010; González-Ibáñez, Muresan, and Wacholder 2011; Riloff et al. 2013; Poria et al. 2016; Zhang, Zhang, and Fu 2016) purely focus on the textual modality and the intra-modal incongruity. Nevertheless, with the advances of multimedia, people tend to express their opinions by multi-modal posts (*e.g.*, text and image) (Lu et al. 2019; Cui et al. 2019; Wei et al. 2019a; Sun et al. 2022). In this context, as shown in Figure 1, if the model ignores the visual information, it would miss the inter-modal incongruity and cannot detect the sarcasm in the social post. Therefore, recent research interests have been drawn to the task of multi-modal sarcasm detection, whose key is to accurately detect the inter- and intra-modal incongruities within the same context.

Figure 1: A multi-modal sarcasm sample from the public dataset (Cai, Cai, and Wan 2019).

In existing works, some models focus on exploiting the whole image feature for incongruity learning (Cai, Cai, and Wan 2019; Xu, Zeng, and Mao 2020; Pan et al. 2020; Liang et al. 2021). Despite their significant progress, they ignore the fact that only some specific visual objects are relevant to the textual context, and the information simply extracted from the whole image is too general. Motivated by this, the recent work (Liang et al. 2022) extracts the object-level feature rather than the global feature of the image, and proposes a cross-modal graph convolutional network that links the visual objects with textual tokens by word similarity to utilize the semantic relations for the multi-modal sarcasm detection. Although this method has achieved promising performance, it suffers from two key limitations. 1) It only emphasizes the visual information in the detected object regions (*i.e.* the area in bounding boxes of Figure 1), but overlooks the contextual relationships between object regions, which also benefit the sarcasm detection. For example, as shown in Figure 1, if we can capture the relationship between the *"man"* and *"woman"* in the *"bus"*, we can infer that *"the bus is crowded"*, and find the sarcasm by referring to the text description. 2) It highly depends on the pre-trained object detection model, and thus is prone to missing objects whose categories are not pre-defined in the object detection model. In light of this, it is inappropriate to abandon the global image-level features, which can complement the object-level feature more or less.

*Corresponding author.

To address these limitations, as shown in Figure 2, we propose a Mutual-enhanced Incongruity Learning Network for multi-modal sarcasm detection, named MILNet, which aims to fully exploit the object-level and global image-level features. Specifically, we design MILNet with four key modules: *multi-modal feature encoding*, *local semantic-guided incongruity learning*, *global incongruity learning*, and *mutual enhancement*. The first module utilizes RoBERTa (Liu et al. 2019) to encode the text, while Faster-RCNN (Anderson et al. 2018) and pre-trained Vision Transformer (ViT) (Dosovitskiy et al. 2021) to encode the image into the image-level and object-level features, respectively. The second module devotes to fully utilizing the inter- and intra-modal semantic relations among objects and text tokens and fulfilling the local semantic-guided incongruity learning, where three graphs (i.e., the text-modal, image-modal, and cross-modal graphs) are built and graph convolutional networks are employed for incongruity learning. Beyond the previous work (Liang et al. 2022), we emphasize the spatial correlation between visual objects by the intersection over union (IoU) scores since it can reflect the relations of objects. Meanwhile, we use a knowledge graph which contains comprehensive object-level semantic relationships to connect visual objects to textual tokens rather than using the conventional lexical similarity that only measures the word-level similarity. The third module targets capturing the incongruity from the global perspective by fusing the text feature and the image-level feature with the multi-head attention mechanism. Ultimately, the fourth module aims to conduct knowledge transferring between the two incongruity learning modules, which should share certain consistency regarding the detection results to boost the performance. Notably, we introduce a sample screening mechanism to ensure the correctness of the transferred knowledge.

Our main contributions can be summarized as follows.

- To the best of our knowledge, we are the first to propose a jointly model for multi-modal sarcasm detection from both local semantic-guided and global aspects by utilizing mutual enhancement.
- We introduce a novel local semantic-guided incongruity learning module by exploiting IoU scores and knowledge graph to extract relations. In addition, we propose a simple but effective global incongruity learning module based on the attention mechanism.
- Extensive experiments demonstrate the superiority of our model over cutting-edge methods. We released the codes and parameters to facilitate the research community[1].

Related Work

Multi-modal Sarcasm Detection. Early works usually make efforts to sarcasm detection only based on text-modal, while with the development of multi-media platforms, multi-modal sarcasm detection has gained a rapid proliferation of interests recently. Schifanella et al. (2016) firstly utilized both textual and visual information to tackle multi-modal sarcasm detection. Cai, Cai, and Wan (2019) created a new dataset from tweet and designed a hierarchical fusion model for the task. Thereafter, Xu, Zeng, and Mao (2020) and Pan et al. (2020) proposed the key of sarcasm detection is capturing incongruities inter-intra-modal. They constructed a decomposition and relation network and a BERT-based model to capture the contradiction, respectively. Liang et al. (2021) realized that sarcastic information is included in some regions of image and some phases in text, and exploited a graph model for drawing incongruous relations between text and image modalities. But this work still utilizes features from the whole image, and is limited by the irrelevant information like all global detection methods above. To alleviate this limitation, Liang et al. (2022) explored a local semantic-guided detection method that they built a cross-modal graph based on visual objects and textual tokens. However, this method heavily depends on the object-extracted technique and misses contextual information. In light of this, we devise a novel model to unify these methods, capturing the underlying consistency between them to boost the performance.

Mutual Learning. Knowledge distillation (Hinton, Vinyals, and Dean 2015; Song et al. 2017; Wen et al. 2021), is a widely used and efficient technique to transfer knowledge from a teacher network to a student network, first introduced by Hinton, Vinyals, and Dean (2015). And then Zhang et al. (2018) extended the form of knowledge distillation and designed mutual learning, which no longer requires a good teacher model, but encourages an ensemble of students to learn collaboratively and teach each other throughout the training process. Due to the fact that mutual learning has achieved better results than traditional distillation, it has gained the attention of many researchers. For example, Wen et al. (2021) resorted to mutual learning to compose the multi-modal query to retrieve the target image. In this work, we propose a mutual enhancement module to encourage consistency between the local semantic-guided incongruity learning module and the global incongruity learning module. In addition, to ensure the correctness of the transferred knowledge, we design a sample screening mechanism.

Methodology

In this section, we first formulate the research problem and then detail the proposed MILNet illustrated in Figure 2.

Problem Formulation

Suppose that we have a set of N training samples $\mathcal{D} = \{s^1, s^2, \cdots, s^N\}$, where each samples $s^i = (\mathcal{T}^i, I^i, Y^i)$ involves three elements. Thereinto, $\mathcal{T}^i = \{t_1^i, t_2^i, \cdots, t_{n_g^i}^i\}$ and I^i denote the textual sentence and image of the i-th sample, respectively, where t_j^i refers to the j-th token of $\mathcal{T}^i$ and n_g^i is the total number of tokens in the sentence $\mathcal{T}^i$. Y^i is the ground truth label of the i-th sample, where $Y^i = 1$ if the sample is sarcastic, and $Y^i = 0$ otherwise. Notably, images may contain embedded text, which usually conveys vital information for capturing the sacarsm. Therefore, we use the optical character recognition text (OCR-text) extracted by Pan et al. (2020), *i.e.*, $\mathcal{O}^i = \{o_1^i, o_2^i, \cdots, o_{n_o^i}^i\}$,

[1]https://frd1228.wixsite.com/milnet.

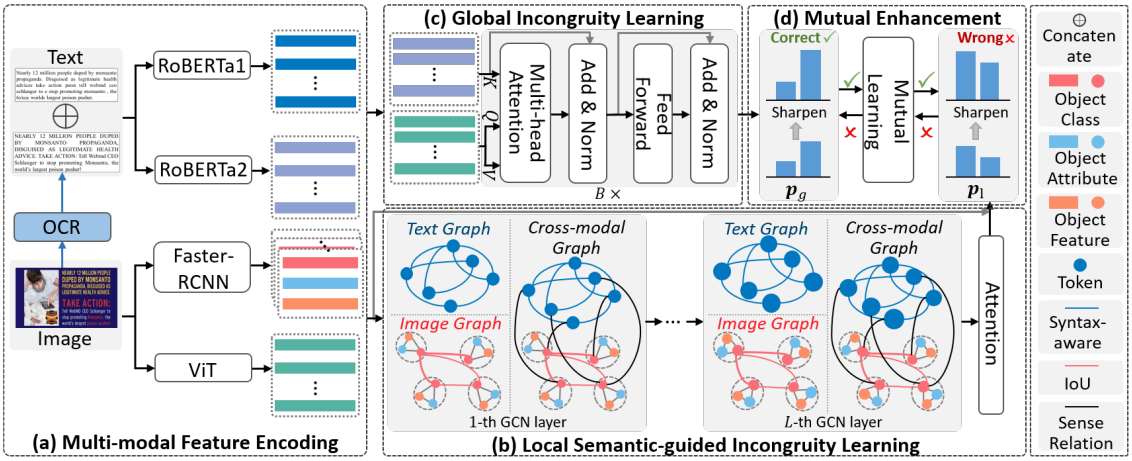


Figure 2: The proposed MILNet consists of four key modules: (a) multi-modal feature encoding, (b) local semantic-guided incongruity learning (LIL), (c) global incongruity learning (GIL), and (d) mutual enhancement.

where o_j^i denotes the j-token of $\mathcal{O}^i$ and n_o^i is the total number of tokens in $\mathcal{O}^i$. In a sense, we aim to devise a novel multi-modal sarcasm detection model $\mathcal{F}$ which can precisely identify whether a given text and its attached image deliver the sarcasm as follows,

$$\mathcal{F}(\mathcal{T}^i,\ I^i,\ \mathcal{O}^i \mid \Theta) \rightarrow \hat{Y}^i, \tag{1}$$

where Θ denotes all the parameters of $\mathcal{F}$, $\hat{Y}^i$ is the binary classification prediction result of the model $\mathcal{F}$. We temporally omit the superscript i that indexes the training samples.

MILNet

As shown in Figure 2, the MILNet comprises four vital components: multi-modal feature encoding, local semantic-guided incongruity learning, global incongruity learning, and mutual enhancement.

Multi-modal Feature Encoding. In this work, multi-modal samples involve two modalities: the text and the image.

Text Encoding. To thoroughly capture the underlying message of the text in multi-modal samples, we concatenate the OCR-text $\mathcal{O}$ with the original text $\mathcal{T}$ and feed them into RoBERTa, which has achieved compelling success in many textual tasks (Dai et al. 2021; Wang et al. 2020), as follows,

$$\boldsymbol{H}^t = \left[\boldsymbol{h}_1^t, \boldsymbol{h}_2^t, \cdots, \boldsymbol{h}_n^t\right] = RoBERTa(\mathcal{T} \oplus \mathcal{O}), \tag{2}$$

where $\boldsymbol{h}_j^t \in \mathbb{R}^{d_h}$ denotes the hidden state vector of j-token, d_h denotes the dimension of the hidden representations, $n = n_g + n_o$ is the total number of tokens after merging the original text and OCR-text, and $\oplus$ refers to the concatenation operation. $\boldsymbol{H}^t$ is the encoded text representation for the input original text and OCR-text.

Image Encoding. To facilitate the following local semantic-guided and global incongruity learning, we extract both object-level and image-level visual features.

Regarding object-level feature extraction, we resort to the Faster-RCNN, which can encode the input image into a set of regional features. To ensure the quality of extracted features, we only select the top k regions with the highest confidence for object-level feature extraction. For each region, we can obtain a visual feature $\boldsymbol{v}_j \in \mathbb{R}^{d_v}$, a positional feature $\boldsymbol{p}_j \in \mathbb{R}^{d_p}$, an object class e_j (*i.e.*,"*bus*" in Figure 1) and an object attribute a_j (*i.e.*,"*white*" in Figure 1) by Faster-RCNN. d_v and d_p denote the dimension of visual and positional features, respectively. As the visual and positional features characterize the object coherently, we fuse them by the linear projections to enrich the visual feature as follows,

$$\boldsymbol{h}_j^f = \boldsymbol{W}_f(\boldsymbol{W}_v\boldsymbol{v}_j + \boldsymbol{p}_j\boldsymbol{W}_p) + \boldsymbol{b}_f, \tag{3}$$

where $\boldsymbol{h}_j^f \in \mathbb{R}^{d_h}$ is the final visual representation of the j-th region. $\boldsymbol{W}_v \in \mathbb{R}^{d_v \times d_v}$, $\boldsymbol{W}_p \in \mathbb{R}^{d_p \times d_v}$, $\boldsymbol{W}_f \in \mathbb{R}^{d_h \times d_v}$ and $\boldsymbol{b}_f \in \mathbb{R}^{d_h}$ are trainable parameters. In addition, the object class e_j and the object attribute a_j are transformed into vectors $\boldsymbol{h}_j^e \in \mathbb{R}^{d_h}$ and $\boldsymbol{h}_j^a \in \mathbb{R}^{d_h}$ through RoBERTa according to Eqn.(2). The image can be finally represented as follows,

$$\begin{cases} \boldsymbol{H}_o^v = [\boldsymbol{I}_1,\ \boldsymbol{I}_2, \ldots,\ \boldsymbol{I}_k], \\ \boldsymbol{I}_j = \left[\boldsymbol{h}_j^e,\ \boldsymbol{h}_j^a, \boldsymbol{h}_j^f\right]^\top, \end{cases} \tag{4}$$

where $\boldsymbol{I}_j \in \mathbb{R}^{3 \times d_h}$ and $\boldsymbol{H}_o^v \in \mathbb{R}^{3k \times d_h}$ are the representations of the j-th region and the final object-level representation of the input image, respectively.

Regarding the image-level features, we adopt the widely-used ViT as the encoder, which splits the input image into r non-overlapping patches and yields r patch embeddings as well as a global embedding corresponding to the special $[CLS]$ token as follows,

$$\boldsymbol{H}_m^v = [\boldsymbol{h}_{cls}^m, \boldsymbol{h}_1^m, \boldsymbol{h}_2^m, \cdots, \boldsymbol{h}_r^m] = ViT\_PE(I), \tag{5}$$

where ViT_PE is the patch embedding layer, $\boldsymbol{h}_j^m \in \mathbb{R}^{d_h}$ represents the j-th patch embedding, $\boldsymbol{h}_{cls} \in \mathbb{R}^{d_h}$ denotes the embedding of the $[CLS]$ token, and $\boldsymbol{H}_m^v \in \mathbb{R}^{(r+1) \times d_h}$ refers to the final image-level representation.

Local Semantic-guided Incongruity Learning (LIL). In fact, there are rich semantic relations among the given multi-modal input. For example, tokens in the textual sentence have semantic associations; objects in the given image have spatial correlations; and moreover, there can be semantic correspondence between tokens in the sentence and objects in the image. These semantic relations undoubtedly would benefit the sarcasm reasoning. In light of this, we build a text-modal graph, an image-modal graph and a cross-modal graph, where edges reflect the above intra- /inter-modal relationships.

Text-modal graph. To capture the semantic relationships existing in textual modality, we build a text-modal graph $\mathcal{G}^t$, whose set of nodes can be denoted as $\{v_1^t, v_2^t, \cdots, v_n^t\}$, where v_j^t corresponds to the j-th token in the given text, and is initialized by the hidden representation extracted by RoBERTa for the j-th token (*i.e.*, $\boldsymbol{h}_j^t$), and the initial representations of the nodes can be defined as $\boldsymbol{G}_1^t = \boldsymbol{H}^t$. Inspired by previous methods (Liang et al. 2021, 2022), the edges of graph are defined by the dependency tree[2], which can reflect the semantic relations among tokens and benefit the textual internal logic reasoning. Concretely, the text-modal adjacency matrix $\boldsymbol{A}^t \in \mathbb{R}^{n\times n}$ is constructed as follows,

$$A_{ij}^t = \begin{cases} 1, & if\ \mathcal{D}(t_i, t_j), i, j \in [1, n] \\ 0, & otherwise, \end{cases} \tag{6}$$

where $\mathcal{D}(t_i, t_j)$ indicates that token t_i (corresponding to the node $\boldsymbol{v}_i^t$) and token t_j (corresponding to the node $\boldsymbol{v}_j^t$) have certain dependency relationships in the dependency tree. To enrich the dependency information of the text, we construct the graph as an undirected graph, which means $A_{ij}^t = A_{ji}^t$. Meanwhile, we set a self-loop for each token (*i.e.*, $A_{ii}^t = 1$). Notably, as the OCR-text can be not a complete sentence, for which the extracted dependency relations can be unreliable, we do not consider the OCR-text in the image for mining the token relationship.

Image-modal graph. To model the semantic relations existing in image modality, we build an image-modal graph $\mathcal{G}^v$, which has $3k$ nodes, denoted as $\{v_1^v, v_2^v, v_3^v, \cdots, v_{3k-2}^e, v_{3k-1}^v, v_{3k}^v\}$, and their initial representations are defined as $\boldsymbol{G}_1^v = \{\boldsymbol{h}_1^e, \boldsymbol{h}_1^a, \boldsymbol{h}_1^f, \cdots, \boldsymbol{h}_k^e, \boldsymbol{h}_k^a, \boldsymbol{h}_k^f\}$. Namely, for each object region, we have three nodes, corresponding to the three feature vectors of the region. As the three vectors essentially describe the same region from different aspects, we fully connect them. Since the spatial relationship between two objects is likely to reflect their semantic correlation, we use the intersection over union (IoU) scores between two object regions to represent their correlations. As the three nodes of each object are mutually connected, we simply select the node referring to the object class as the representative node for linking the different object regions. Mathematically, the image-modal adjacency matrix $\boldsymbol{A}^v \in \mathbb{R}^{3k\times 3k}$ can be summarized as follows,

$$A_{ij}^v = \begin{cases} 1, & if\ i\ mod\ 3 = j\ mod\ 3, \\ S_{i,j}, & if\ i\ rem\ 3 = 1, j\ rem\ 3 = 1, \\ 0, & otherwise, \end{cases} \tag{7}$$

where mod denotes the modulo operation, which helps link the three nodes for the same object. rem denotes the remainder operation, used for linking different object regions. $S_{i,j}$ is the IoU score between $(i\ mod\ 3)$-th and $(j\ mod\ 3)$-th object regions, $i, j \in [1, 3k]$.

Cross-modal graph. To learn semantic relations across different modalities, we construct a cross-modal graph $\mathcal{G}^c$, whose nodes of this graph cover all text token and visual object representations, denoted as $\{v_1^c, \cdots, v_{n+3k}^c\}$, and their initial representations are defined as $\boldsymbol{G}_1^c = \{\boldsymbol{h}_1^t, \cdots, \boldsymbol{h}_n^t, \boldsymbol{h}_1^e, \boldsymbol{h}_1^a, \boldsymbol{h}_1^f, \cdots, \boldsymbol{h}_k^e, \boldsymbol{h}_k^a, \boldsymbol{h}_k^f\}$). To build the cross-modal edges, we resort to the knowledge graph ConceptNet5[3], to obtain the semantic relationships between textual tokens and object classes/attributes. Specifically, if a textual token has certain relation with a visual object's class/attribute label according to the ConceptNet5, we will link their corresponding nodes with the weight of 1. Notably, to enrich the information contained in the cross-modal graph, we also integrate the edges of the two single-modal graphs (i.e., $\mathbf{A}^t$ and $\mathbf{A}^v$) into it (Liang et al. 2021, 2022). Ultimately, the cross-modal adjacency matrix $\boldsymbol{A}^c \in \mathbb{R}^{(n+3k)\times(n+3k)}$ for each input multi-modal sample can be formulated as follows,

$$A_{ij}^c = \begin{cases} A_{ij}^t, & if\ A_{ij}^t > 0,\ i, j \in [1, n], \\ A_{ij}^v, & if\ A_{ij}^v > 0,\ i, j \in [n+1, n+3k], \\ 1, & if\ K(t_i, e_{(j-n-1)/3+1}) \\ & or\ K(t_i, a_{(j-n-2)/3+1})), \\ 0, & otherwise, \end{cases} \tag{8}$$

where A_{ij}^t and A_{ij}^v are the integrated edges in text-/image-modal graphs, respectively. $K(t_i, e_{(j-n-1)/3+1})$ indicates that the token t_i and the object class $e_{(j-n-1)/3+1}$ have certain relation in the ConceptNet5, where $i \in [1, n_g]$, $j \in \{j \in [n+1, n+3k] | (j-n) rem 3 = 1\}$. Similarly, $K(t_i, a_{(j-n-2)/3+1})$ denotes that the token t_i and the object attribute have certain relation in the ConceptNet5, where $i \in [1, n_g]$, $j \in \{j \in [n+1, n+3k] | (j-n) rem 3 = 2\}$.

Graph Convolutional Network. Thereafter, we resort to the graph convolution networks (GCNs) (Kipf and Welling 2017) to mine the above-defined relationship and learn the multi-modal incongruity. GCNs work on updating node features with their neighborhoods according to the adjacency matrix and show their superiority on several tasks (Wei et al. 2019b; Jing et al. 2022; Zhuang and Hasan 2022; Wang et al. 2022a; He et al. 2022; Wang et al. 2022b). Specifically, we utilize the GCNs to iteratively learn the intra-modal incongruities and the inter-modal incongruities. The process is defined as follows,

$$\begin{cases} \boldsymbol{G}_{u'}^t = ReLU(\tilde{\boldsymbol{A}}^t \boldsymbol{G}_{u-1}^t \boldsymbol{W}_u^t + \boldsymbol{b}_u^t), \\ \boldsymbol{G}_{u'}^v = ReLU(\tilde{\boldsymbol{A}}^v \boldsymbol{G}_{u-1}^v \boldsymbol{W}_u^v + \boldsymbol{b}_u^v), \\ \boldsymbol{G}_u^c = \boldsymbol{G}_u^t \oplus \boldsymbol{G}_u^v = ReLU(\tilde{\boldsymbol{A}}^c (\boldsymbol{G}_{u'}^t \oplus \boldsymbol{G}_{u'}^v) \boldsymbol{W}_u^c + \boldsymbol{b}_u^c), \end{cases} \tag{9}$$

where $\tilde{\boldsymbol{A}}^x = (\boldsymbol{D}^x)^{-\frac{1}{2}} \boldsymbol{A}^x (\boldsymbol{D}^x)^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, and $\boldsymbol{D}^x$ is the degree matrix of $\boldsymbol{A}^x$. $\boldsymbol{G}_u^x$ are the representations of nodes in corresponding graphs after the u-th GCN process, where $x \in \{t, v, c\}$, $u \in [1, U]$,

[2]https://spacy.io/.

[3]https://github.com/commonsense/conceptnet5.

and U denotes the total number of iterations. In addition, $\{\boldsymbol{W}_l^t, \boldsymbol{W}_l^v, \boldsymbol{W}_l^c\} \in \mathbb{R}^{d_h \times d_h}$ and $\{\boldsymbol{b}_l^t, \boldsymbol{b}_l^v, \boldsymbol{b}_l^c\} \in \mathbb{R}^{d_h}$ are trainable parameters of the l-th GCN layer.

Subsequently, following (Liang et al. 2022; Zhang, Li, and Song 2019), we utilize a retrieval-based attention mechanism to obtain a graph-oriented cross-modal representation for detection. The intention is to retrieve the crucial representation in inter-/intra-graph. Specifically, we feed the initial node representations of cross-modal graph (*i.e.*, $\boldsymbol{H} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{n+3k}\}$) and the final outputs of the GCN layers (*i.e.*, $\boldsymbol{G}_L^c = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_{n+3k}\}$) into attention mechanism. The process is defined by following transformation,

$$\begin{cases} \boldsymbol{\alpha}_i = softmax(\sum_{j=1}^{n+3k} \boldsymbol{v}_j^\top \boldsymbol{g}_j), \\ \boldsymbol{f}_l = \sum_{i=1}^{n+3k} \boldsymbol{\alpha}_i \boldsymbol{h}_i, \end{cases} \tag{10}$$

where $\boldsymbol{\alpha}_i$ refers to attention scores, $\boldsymbol{f}_l$ is the final sarcasm representation of LIL for sarcasm detection. And then we feed it into fully connected layers to gain the predicted probability distributions as follows,

$$\boldsymbol{p}_l = softmax(\boldsymbol{W}_l \boldsymbol{f}_l + \boldsymbol{b}_l), \tag{11}$$

where $\boldsymbol{p}_l \in \mathbb{R}^2$ is the predicted probability vector of LIL, $\boldsymbol{W}_l \in \mathbb{R}^{d \times 2}$, $\boldsymbol{b}_l \in \mathbb{R}^2$ are trainable parameters. Ultimately, we calculate the cross-entropy loss to supervise our LIL module as follows,

$$\mathcal{L}_{ce}^l = y^i log p_l^i + (1 - y^i) log(1 - p_l^i) + \lambda_1 \parallel \Theta_l \parallel^2, \tag{12}$$

where Θ_l denotes all trainable parameters in the LIL module. y^i and p_l^i are the i-th elements of the ground truth $\boldsymbol{y}$ and $\boldsymbol{p}_l$, respectively. λ_1 represents the weight coefficients of the Frobenius norm to prevent modules from overfitting.

Global Incongruity Learning (GIL). As for GIL, to enforce the module focus on the input image-level features that are highly correlated with sarcasm, we construct it based on the widely-used attention mechanism (Vaswani et al. 2017). In particular, we use image-level feature $\boldsymbol{H}_m^v$ defined in Eqn.(5) as original query, the text feature $\boldsymbol{H}^t$ defined in Eqn.(2) as key and value. In this way, the image features guide our model to pay more attention to the incongruous text phases by the attention mechanism layer follows,

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{H}_m^v \boldsymbol{W}^Q, \boldsymbol{K} = \boldsymbol{H}^t \boldsymbol{W}^K, \boldsymbol{V} = \boldsymbol{H}^t \boldsymbol{W}^V \\ \boldsymbol{H}' = softmax(\frac{\boldsymbol{K}\boldsymbol{Q}^\top}{\sqrt{d_h}} \boldsymbol{V}), \\ \hat{\boldsymbol{H}} = LN(\boldsymbol{H}' + \boldsymbol{H}_m^v), \\ \check{\boldsymbol{H}} = LN(\hat{\boldsymbol{H}} + (\boldsymbol{W}^M \hat{\boldsymbol{H}} + \boldsymbol{b}^M)), \end{cases} \tag{13}$$

where the query $\boldsymbol{Q} \in \mathbb{R}^{(r+1) \times d_h}$ is projected from the encoded visual feature $\mathbf{H}_m^v$, the key $\boldsymbol{K} \in \mathbb{R}^{n \times d_h}$ and the value $\boldsymbol{V} \in \mathbb{R}^{n \times d_h}$ are both projected from the encoded textual feature $\mathbf{H}^t$. $LN(\cdot)$ refers to layer normalization operation (Ba, Kiros, and Hinton 2016), and $\boldsymbol{H}' \in \mathbb{R}^{(r+1) \times d_h}$ is the output of attention mechanism, $\hat{\boldsymbol{H}} \in \mathbb{R}^{(r+1) \times d_h}$ represents the hidden representations of the residual connection, and $\check{\boldsymbol{H}} \in \mathbb{R}^{(r+1) \times d_h}$ which includes textual and visual information becomes the new queries to lead the model to extract important information. In addition, $\{\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{W}^M\} \in \mathbb{R}^{d_h \times d_h}$ and $\boldsymbol{b}^M \in \mathbb{R}_h^d$ are to-be-learned parameters. We stack B such attention mechanism layers, get the output of the last layer as $\check{\boldsymbol{H}}_B$ and treat encoding of [CLS] token in the output $\check{\boldsymbol{H}}_B$ as the final representation $\boldsymbol{f}_g$. Similar to the LIL, we utilize fully connected layers to obtain the predicted probability distributions of GIL $\boldsymbol{p}_g \in \mathbb{R}^2$ as follows,

$$\boldsymbol{p}_g = softmax(\boldsymbol{W}_g \boldsymbol{f}_g + \boldsymbol{b}_g), \tag{14}$$

where $\boldsymbol{W}_g \in \mathbb{R}^{d \times 2}$, $\boldsymbol{b}_g \in \mathbb{R}^2$ are to-be-learned parameters and the cross-entropy loss can be defined as follows,

$$\mathcal{L}_{ce}^g = y^i log p_l^i + (1 - y^i) log(1 - p_l^i) + \lambda_2 \parallel \Theta_g \parallel^2, \tag{15}$$

where y^i and p_g^i are the i-th elements of the ground truth $\boldsymbol{y}$ and $\boldsymbol{p}_g$, respectively.Θ_g denotes all trainable parameters in GIL module and λ_2 denotes the weight coefficients.

Mutual Enhancement. As both local semantic-guided and global incongruity learning modules aim to capture the inter- and intra-modality incongruities, there should be certain intrinsic consistency between the two modules. In view of this, we make the two modules share knowledge with each other by adopting mutual learning (Zhang et al. 2018). Specifically, we employ the Kullback Leibler (KL) Divergence between $\boldsymbol{p}_l$ and $\boldsymbol{p}_g$, which can measure the differences between two distributions to encourage consistency between the two learning modules. To avoid incorrect knowledge being transferred, different from the existing methods that transfer knowledge regardless of samples, we propose to only transfer reliable knowledge. In particular, we introduce an indicator controlling whether to transfer the prediction result of this sample. Formally, the objective function for knowledge transferring can be written as follows,

$$\begin{cases} \mathcal{L}_{kl}^{g \to l} = \eta_1 D_{KL}(\boldsymbol{p}_g \parallel \boldsymbol{p}_l), \\ \mathcal{L}_{kl}^{l \to g} = \eta_2 D_{KL}(\boldsymbol{p}_l \parallel \boldsymbol{p}_g), \end{cases} \tag{16}$$

where $(g \to l)$ denotes the knowledge transferring from the GIL to LIL, and similarly, $(l \to g)$ denotes the knowledge transferring from the LIL to GIL. $\eta_{1/2}$ are the control parameters to avoid incorrect knowledge transferring which is defined as follows,

$$\eta_{1/2} = \begin{cases} 1, if\ argmax(\boldsymbol{p}_{g/l}) = Y \\ 0, otherwise \end{cases} \tag{17}$$

where $argmax$ denotes the operation that gains the predicted labels from the predicted result $\boldsymbol{p}_{g/l}$, and $\boldsymbol{p}_{g/l}$ is the predicted probability distribution of the mudule GIL/LIL which shares the knowledge (*i.e.*, $\boldsymbol{p}_g$ for $\mathcal{L}_{kl}^{g \to l}$ and $\boldsymbol{p}_l$ for $\mathcal{L}_{kl}^{l \to g}$). Similar to (Hinton, Vinyals, and Dean 2015), we also sharpen the predicted distribution of the model with a temperature parameter τ for knowledge transfer.

We combine all loss functions as follows,

$$\begin{cases} \mathcal{L}^l = \mathcal{L}_{ce}^l + \delta_1 \mathcal{L}_{kl}^{g \to l}, \\ \mathcal{L}^g = \mathcal{L}_{ce}^g + \delta_2 \mathcal{L}_{kl}^{l \to g}, \end{cases} \tag{18}$$

MODALITY	METHOD	Acc (%)	F1-score			Macro-average		
			Pre (%)	Rec (%)	F1 (%)	Pre (%)	Rec (%)	F1 (%)
single-modal	Image (Cai, Cai, and Wan 2019)†‡	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT (Dosovitskiy et al. 2021)‡	73.72	65.56	71.64	68.46	72.78	73.37	72.97
	TextCNN (Kim 2014)†∗	80.03	74.29	76.39	75.32	78.03	78.28	78.15
	Bi-LSTM†∗	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN (Tay et al. 2018)†∗	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD (Xiong et al. 2019)†∗	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT (Devlin et al. 2019)†∗	83.85	78.72	82.27	80.22	81.31	80.87	81.09
	RoBERTa (Liu et al. 2019)∗	85.51	78.24	88.11	82.88	84.83	85.95	85.16
multi-modal	HFM (Cai, Cai, and Wan 2019) †	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	D&R Net (Xu, Zeng, and Mao 2020) †	84.02	77.97	83.42	80.60	-	-	-
	Res-BERT (Pan et al. 2020) †	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT (Pan et al. 2020) †	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs (Liang et al. 2021) †	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	CMGCN (Liang et al. 2022) †	<u>87.55</u>	<u>83.63</u>	<u>84.69</u>	<u>84.16</u>	<u>87.02</u>	<u>86.97</u>	<u>87.00</u>
	MILNet	**89.50**	**85.16**	**89.16**	**87.11**	**88.88**	**89.44**	**89.12**

Table 1: Performance comparison among different methods on the multi-modal sarcasm dataset in terms of Acc, F1-score and Macro-average F1-score. † indicates the results are cited from (Liang et al. 2022). ‡ denotes models only utilize visual information as input and ∗ denotes modes only utilize textual information as input. The best results are highlighted in boldface, while the second-best results are underlined.

where δ_1 and δ_2 are non-negative hyper-parameters. Ultimately, the binary classification prediction result $\hat{Y}$ is defined as follows,

$$\hat{Y} = argmax(\frac{\boldsymbol{p}_g + \boldsymbol{p}_l}{2}). \tag{19}$$

We calculated the cross-entropy loss of the validating set and reserved the model with the best performance for testing.

Experiment

Experimental Settings

Dataset. Following previous works, we evaluated our model on a public available multi-modal sarcasm detection dataset (Cai, Cai, and Wan 2019) with English tweets. Thereinto, tweets with some special hashtags (*e.g.* sarcasm) are positive examples and those without such hashtags are negative examples. Furthermore, the dataset is divided into a training set, a validating set, and a testing set, which includes $19,816$, $2,410$, and $2,409$ samples, respectively. In addition, we refer readers to the supplementary material for the implementation details.

On Model Comparison (RQ1)

To validate the effectiveness of our MILNet, we compared it with several state-of-art baselines which can be broadly categorized into two groups. (1) *Single-modal Methods.* These methods simply take visual or textual information as input for multi-modal sarcasm detection, including: **Image** (Cai, Cai, and Wan 2019); **ResNet-based** (He et al. 2016); **ViT** (Dosovitskiy et al. 2021); **TEXTCNN**; (Kim 2014); **Bi-LSTM**; **SIARN** (Tay et al. 2018); **SMSD** (Xiong et al. 2019); **BERT** (Devlin et al. 2019) and **RoBERTa** (Liu et al. 2019). (2) *Multi-modal Methods.* These methods exploit both visual and textual information as input for multi-modal sarcasm detection, including: **HFM** (Cai, Cai, and Wan 2019); **D&R Net** (Xu, Zeng, and Mao 2020); **Res-BERT** (Pan et al. 2020); **Att-BERT** (Pan et al. 2020); **InCross-MGs** (Liang et al. 2021) and **CMGCN** (Liang et al. 2022).

Table 1 illustrates the performance comparison among different methods. From this table, we had the following observations. 1) MILNet consistently outperforms both single-modal and multi-modal baselines across different evaluation metrics, which denotes that MILNet can significantly improve the performance of sarcasm detection over state-of-the-art methods. 2) Our model surpasses both InCross-MGs (global) and CMGCN (local semantic-guided). This implies the advantage of incorporating both global and local semantic-guided methods and demonstrates these two kinds of methods can complement each other. 3) Multi-modal methods perform better than single-modal baselines overall, which indicates that simultaneously extracting incongruities from textual and visual information can improve the performance of sarcasm detection. And 4) to justify the improvement is statistically significant, we also conducted the significant test between our results and the second best results, and found that all the p-values are less than 0.01. This validates the superiority of MILNET over existing methods.

On Ablation Study (RQ2)

To explore the roles of different components in our proposed model, we compared MILNet with the following derivations. 1) **LIL-only** and **GIL-only**. To explore the effect of both local semantic-guided and global incongruity learning modules, we removed the local semantic-guided incongruity learning module and the global incongruity learning module, respectively. 2) **w/o-mutual-learning**. To validate the necessity of the mutual enhancement, we omitted the knowledge distillation between the LIL and GIL modules by directly averaging their output features. 3) **w/o-sample-screening**. We disabled the sample screening in the mutual enhancement to get more insight into it. 4) **w/o-embedded-text**. To verify the importance of the OCR-text extracted from image, we discarded it and only

MODEL	Acc. (%)	F1 (%)	Macro-F1 (%)
LIL-only	88.00	84.70	87.41
GIL-only	88.25	85.49	87.81
w/o-embedded-text	89.29	86.67	88.86
w/o-mutual-learning	88.95	86.62	88.61
w/o-sample-screening	88.96	86.57	88.60
w/o-text-modal-relations	88.05	85.80	87.74
w/o-image-modal-relations	89.12	86.85	88.79
w/o-cross-modal-relations	88.96	86.65	88.62
w/-similarity	89.00	86.81	88.69
MILNet	**89.50**	**87.11**	**89.12**

Table 2: Experiment results of ablation study.

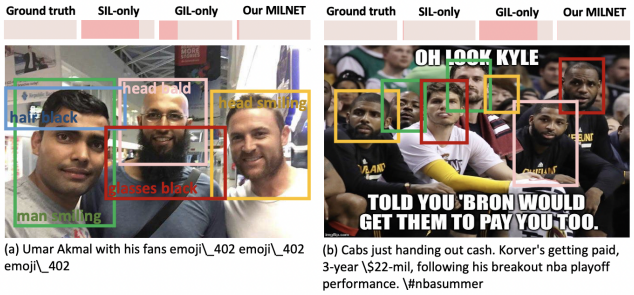
fed the original sentence from text modality into our MILNet framework. 5) **w/o-text-modal-relations**, **w/o-image-modal-relations** and **w/o-cross-modal-relations**. To check the effect of text-modal relations, image-modal relations and cross-modal relations, we removed these relations by replacing the inter-/intra-adjacency matrix with identity matrix, respectively. And 6) **w/-similarity**. To demonstrate the superiority of the semantic relationships in knowledge graphs, we adopted lexical similarity[4] instead of knowledge graph to fulfill cross-modal relations for comparison.

Table 2 summarizes the performance of MILNet with its derivations. From this table, we can draw the following observations. 1) Our MILNet surpasses both LIL-only and GIL-only, demonstrating that removing either the global incongruity learning module or the local semantic-guided incongruity learning module will hurt the performance of MILNet. 2) Both the LIL-only and GIL-only surpass the state-of-art local semantic-guided sarcasm detection baselines(*i.e.,* CMGCN and InCrossMGs). This suggests that the semantic-guided incongruities and global sarcasm detection incongruities can be well captured by MILNet in an efficient way. 3) MILNet exceeds w/o-embedded-text, denoting that the text embedded in image does evolve vital semantic for multi-modal sarcasm detection. 4) MILNet outperforms w/o-mutual-learning. This verifies the knowledge does share between LIL and GIL modules, and the advantage of integrating the two compositions via the mutual learning. 5) Some of the indicators in w/o-sample-screening are even lower than those in w/o-mutual-learning, indicating that choosing the right knowledge to transfer is necessary. 6) All of the w/o-text-modal-relations, w/o-image-modal-relations, w/o-cross-modal-relations perform worse than MILNet, demonstrating that inter/intra-modalities relations do contribute to the multi-modal sarcasm detection, which can be well captured by the proposed MILNet. And 7) w/-similarity exceeds w/o-cross-modal-relations but still performs worse than our MILNet, confirming that the lexical similarity can capture partially effective semantic relationships but is not as comprehensive as knowledge graphs.

On Case Study (RQ3)

To get an intuitive understanding of how the MILNet works on multi-modal sarcasm detection, we exhibited two testing

[4]https://www.nltk.org/.



(a) Umar Akmal with his fans emoji_402 emoji_402 emoji_402

(b) Cabs just handing out cash. Korver's getting paid, 3-year \$22-mil, following his breakout nba playoff performance. \#nbasummer

Figure 3: Example of case study, where (a) is the sample misclassified by LIL-only, (b) is the sample misclassified by GIL-only. The bars on the top of the image represent the predictions of the corresponding models. The dark and light colors denote probabilities of 1 and 0, respectively.

samples in Figure 3. In the case (a), the LIL-only module gets the object classes like *"head"* and *"hair"*, and even finds *the man in the middle is bald* but it still fails to capture the funny relationship between their beards and their hair since the LIL-only overlooks the abstract contextual relations out of the boxes. However, extracting these abstract contextual relations is easy for the GIL-only module and it does work great on this sample. In addition, the *"beard"* in (a) is misclassified as *"grasses"*, which interferes with the model classification and reconfirms that the performance of object extraction techniques would limit the semantic-based incongruity learning module. As for sample (b), we omitted the object classes and object attributes for brief. The GIL-only module is unable to label this example correctly while the LIL-only module does it well, mainly because the LIL-only model concentrates on the people's faces that can easily capture their expression. As shown in the bars, our MILNet achieves excellent results on both samples, which shows that the knowledge of LIL module and GIL module is indeed shared with each other, and our MILNet successfully combines the advantages of the two modules.

Conclusion and Future Work

In this work, we present a mutual-enhanced incongruity learning network for multi-modal sarcasm detection (MILNet), which seamlessly unifies the local semantic-guided incongruity learning module and the global incongruity learning module. Extensive experiments on the public available multi-modal sarcasm detection dataset demonstrate the superiority of our model over state-of-art methods. In particular, we notice that even the LIL-only and GIL-only outperform the state-of-art models, which suggests the way that we construct the two modules is effective. Meanwhile, the ablation study justifies the necessity of mutual enhancement that simultaneously incorporation the two incongruity learning modules and verifies the importance of inter- and intra-relations. To take it further, we will explore more techniques to align different modalities.

Acknowledgments

We want to thank our anonymous reviewers for their feedback. This work is supported by the National Key R&D Program of China (Response-driven intelligent enhanced analysis and control for bulk power system stability, 2021YFB2400800), the National Natural Science Foundation of China (U1936203), the Shandong Provincial Natural Science Foundation (ZR2022YQ59).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086.

Ba, L. J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *CoRR*, abs/1607.06450.

Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *ACL*, 2506–2515.

Cui, H.; Zhu, L.; Li, J.; Yang, Y.; and Nie, L. 2019. Scalable Deep Hashing for Large-scale Social Image Retrieval. *IEEE Transactions on Image Processing*, 29: 1271–1284.

Dai, J.; Yan, H.; Sun, T.; Liu, P.; and Qiu, X. 2021. Does syntax matter? A Strong Baseline for Aspect-based Sentiment Analysis with RoBERTa. In *NAACL-HLT*, 1816–1829.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semi-Supervised Recognition of Sarcasm in Twitter and Amazon. In *CoNLL*, 107–116.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

González-Ibáñez, R. I.; Muresan, S.; and Wacholder, N. 2011. Identifying Sarcasm in Twitter: A Closer Look. In *ACL*, 581–586.

He, D.; Liang, C.; Liu, H.; Wen, M.; Jiao, P.; and Feng, Z. 2022. Block Modeling-Guided Graph Convolutional Neural Networks. In *AAAI*, 4022–4029.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Jing, L.; Tian, M.; Chen, X.; Sun, T.; Guan, W.; and Song, X. 2022. CI-OCM: Counterfactural Inference towards Unbiased Outfit Compatibility Modeling. In *Proceedings of the 1st Workshop on Multimedia Computing towards Fashion Recommendation*, 31–38.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *ACL*, 1746–1751.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Liang, B.; Lou, C.; Li, X.; Gui, L.; Yang, M.; and Xu, R. 2021. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs. In *MM*, 4707–4715.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In *ACL*, 1767–1777.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, X.; Zhu, L.; Cheng, Z.; Nie, L.; and Zhang, H. 2019. Online Multi-modal Hashing with Dynamic Query-adaption. In *SIGIR*, 715–724.

Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; and Wang, W. 2020. Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection. In *Findings of EMNLP*, 1383–1392.

Poria, S.; Cambria, E.; Hazarika, D.; and Vij, P. 2016. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *COLING*, 1601–1612.

Riloff, E.; Qadir, A.; Surve, P.; Silva, L. D.; Gilbert, N.; and Huang, R. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *EMNLP*, 704–714.

Schifanella, R.; de Juan, P.; Tetreault, J. R.; and Cao, L. 2016. Detecting Sarcasm in Multimodal Social Platforms. In *MM*, 1136–1145.

Song, X.; Feng, F.; Liu, J.; Li, Z.; Nie, L.; and Ma, J. 2017. NeuroStylist: Neural Compatibility Modeling for Clothing Matching. In *MM*, 753–761.

Sun, T.; Wang, W.; Jing, L.; Cui, Y.; Song, X.; and Nie, L. 2022. Counterfactual Reasoning for Out-of-distribution Multimodal Sentiment Analysis. In *MM*, 15–23.

Tay, Y.; Luu, A. T.; Hui, S. C.; and Su, J. 2018. Reasoning with Sarcasm by Reading In-Between. In *ACL*, 1010–1020.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, T.; Jin, D.; Wang, R.; He, D.; and Huang, Y. 2022a. Powerful Graph Convolutional Networks with Adaptive Propagation Mechanism for Homophily and Heterophily. In *AAAI*, 4210–4218.

Wang, X.; Sun, X.; Yang, T.; and Wang, H. 2020. Building a bridge: A Method for Image-text Sarcasm Detection without Pretraining on Image-text Data. In *Proceedings of the first international workshop on natural language processing beyond text*, 19–29.

Wang, Y.; Cao, M.; Fan, Z.; and Peng, S. 2022b. Learning to Detect 3D Facial Landmarks via Heatmap Regression with Graph Convolutional Network. In *AAAI*, 2595–2603.

Wei, Y.; Wang, X.; Guan, W.; Nie, L.; Lin, Z.; and Chen, B. 2019a. Neural Multimodal Cooperative Learning toward Micro-video Understanding. *IEEE Transactions on Image Processing*, 29: 1–14.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 2019b. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *MM*, 1437–1445.

Wen, H.; Song, X.; Yang, X.; Zhan, Y.; and Nie, L. 2021. Comprehensive Linguistic-Visual Composition Network for Image Retrieval. In *SIGIR*, 1369–1378.

Xiong, T.; Zhang, P.; Zhu, H.; and Yang, Y. 2019. Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling. In *WWW*, 2115–2124.

Xu, N.; Zeng, Z.; and Mao, W. 2020. Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association. In *ACL*, 3777–3786.

Zhang, C.; Li, Q.; and Song, D. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *EMNLP-IJCNLP*, 4567–4577.

Zhang, M.; Zhang, Y.; and Fu, G. 2016. Tweet Sarcasm Detection Using Deep Neural Network. In *COLING*, 2449–2460.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *CVPR*, 4320–4328.

Zhuang, J.; and Hasan, M. A. 2022. Defending Graph Convolutional Networks against Dynamic Graph Perturbations via Bayesian Self-Supervision. In *AAAI*, 4405–4413.