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} @gmail.com

Abstract

Sarcasm is a sophisticated linguistic phenomenon that aims to express people’s emotions contrary to the normal interpretations, which frequently appears 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 today’s 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.

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.
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.

**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 $\mathcal{D} = \{s^1, s^2, \ldots, s^N\}$, where each samples $s^i = (T^i, I^i, Y^i)$ involves three elements. Thereinto, $T^i = \{t^i_1, t^i_2, \ldots, t^i_{n^i_o}\}$ and $I^i$ denote the textual sentence and image of the $i$-th sample, respectively, where $t^i_j$ refers to the $j$-th token of $T^i$ and $n^i_o$ is the total number of tokens in the sentence $T^i$. $Y^i$ 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 sarcastism. Therefore, we use the optical character recognition text (OCR-text) extracted by Pan et al. (2020), i.e., $O^i = \{o^i_1, o^i_2, \ldots, o^i_{n^i_o}\}$.

where $o_j^i$ denotes the $j$-token of $O^i$ and $n_o^i$ is the total number of tokens in $O^i$. 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,

$$\mathcal{F}(T^i, I^i, O^i | \Theta) \rightarrow \hat{y}^i,$$

where $\Theta$ denotes all the parameters of $\mathcal{F}$, $\hat{y}^i$ is the binary classification prediction result of the model $\mathcal{F}$. We temporarily 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 $O$ with the original text $T$ and feed them into RoBERTa, which has achieved compelling success in many textual tasks (Dai et al. 2021; Wang et al. 2020), as follows,

$$H^t = [h_1^t, h_2^t, \cdots, h_n^t] = \text{RoBERTa}(T \oplus O),$$

where $h_j^t \in \mathbb{R}^{d_h}$ 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. $H^t$ 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 $v_j \in \mathbb{R}^{d_v}$, a positional feature $p_j \in \mathbb{R}^{d_p}$, 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,

$$h_j^f = W_f(W_e v_j + p_j W_p) + b_f,$$

where $h_j^f \in \mathbb{R}^{d_h}$ is the final visual representation of the $j$-th region. $W_v \in \mathbb{R}^{d_v \times d_h}$, $W_p \in \mathbb{R}^{d_p \times d_v}$, $W_f \in \mathbb{R}^{d_h \times d_v}$ and $b_f \in \mathbb{R}^{d_h}$ are trainable parameters. In addition, the object class $e_j$ and the object attribute $a_j$ are transformed into vectors $h_{e_j} \in \mathbb{R}^{d_e}$ and $h_{a_j} \in \mathbb{R}^{d_a}$ through RoBERTa according to Eqn.(2). The image can be finally represented as follows,

$$\begin{align*}
H_v^o &= [I_1, I_2, \ldots, I_k], \\
I_j &= [h_{e_j}^o, h_{a_j}^o, h_j^f]^T,
\end{align*}$$

where $I_j \in \mathbb{R}^{k \times d_h}$ and $H_v^o \in \mathbb{R}^{k \times d_h}$ 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,

$$H_v^m = [h_{cls}^m, h_1^m, h_2^m, \ldots, h_r^m] = \text{ViT}PE(I),$$

where $\text{ViT}PE$ is the patch embedding layer, $h_j^m \in \mathbb{R}^{d_h}$ represents the $j$-th patch embedding, $h_{cls} \in \mathbb{R}^{d_h}$ denotes the embedding of the $[CLS]$ token, and $H_v^m \in \mathbb{R}^{(r+1) \times d_h}$ 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 \( G' \), whose set of nodes can be denoted as \( \{v'_1, v'_2, \cdots, v'_n\} \), where \( v'_i \) corresponds to the \( i \)-th token in the given text, and is initialized by the hidden representation extracted by RoBERTa for the \( i \)-th token (i.e., \( h'_i \)), and the initial representations of the nodes can be defined as \( G'_1 = H' \). 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 \( A^t \in \mathbb{R}^{n \times n} \) is constructed as follows,

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

where \( D(t_i, t_j) \) indicates that token \( t_i \) (corresponding to the node \( v'_i \)) and token \( t_j \) (corresponding to the node \( v'_j \)) 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 \( A^t_{ij} = A^t_{ji} \). Meanwhile, we set a self-loop for each token (i.e., \( A^t_{ii} = 1 \)). 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 \( G'' \), which has 3\( k \) nodes, denoted as \( \{v''_1, v''_2, \cdots, v''_{3k-2}, v''_{3k-1}, v''_{3k}\} \), and their initial representations are defined as \( G''_1 = \{h''_1, h''_2, h''_3, \cdots, h''_{3k-2}, h''_{3k-1}, h''_{3k}\} \). 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 \( A^u \in \mathbb{R}^{3k \times 3k} \) can be summarized as follows,

\[
A^u_{ij} = \begin{cases} 
1, & \text{if } i \mod 3 = j \mod 3, \\
S_{i,j}, & \text{if } i \rem 3 = 1, j \rem 3 = 1, \\
0, & \text{otherwise},
\end{cases} \tag{7}
\]

where \( \text{mod} \) denotes the modulo operation, which helps link the three nodes for the same object. \( \text{rem} \) 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 \( G^c \), whose nodes of this graph cover all text token and visual object representations, denoted as \( \{v'_1, \cdots, v'_{n+3k}\} \), and their initial representations are defined as \( G^c_1 = \{h'_1, \cdots, h'_n, h''_1, h''_2, \cdots, h''_{3k}, h''_{3k+1}\} \). To build the cross-modal edges, we resort to the knowledge graph ConceptNet\(^3\), 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., \( A^t \) and \( A^u \)) into it (Liang et al. 2021, 2022). Ultimately, the cross-modal adjacency matrix \( A^c \in \mathbb{R}^{(n+3k) \times (n+3k)} \) for each input multi-modal sample can be formulated as follows,

\[
A^c_{ij} = \begin{cases} 
A^t_{ij}, & \text{if } A^t_{ij} > 0, i, j \in [1, n], \\
A^u_{ij}, & \text{if } A^u_{ij} > 0, i, j \in [n+1, n+3k], \\
1, & \text{if } K(t_i, a_{(j-(n+1)/3+1)}) \\
0, & \text{otherwise}.
\end{cases} \tag{8}
\]

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

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,

\[
\begin{align*}
G'_v &= \text{ReLU}(\tilde{A}' v G'_v - W'_v + b'_v), \\
G''_v &= \text{ReLU}(\tilde{A}' v G''_v - W''_v + b''_v), \\
G^c_v &= G'_v \odot G''_v = \text{ReLU}(\tilde{A}' v (G'_v \odot G''_v)W^c_v + b^c_v),
\end{align*}
\tag{9}
\]

where \( \tilde{A}' = (D^+)^{-\frac{1}{2}}A\tilde{D}^+(D^+)^{-\frac{1}{2}} \) is the normalized symmetric adjacency matrix, and \( D^+ \) is the degree matrix of \( A^c \). \( G'_v, G''_v \) 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\text{https://spacy.io/}.
\[^3\text{https://github.com/commonsense/conceptnet5.}\]
and $U$ denotes the total number of iterations. In addition, \( \{ W_i^l, W_l^+, W_l^- \} \in \mathbb{R}^{d_h \times d_h} \) and \( \{ b_i, b_o^l, b_o^c \} \in \mathbb{R}^{d_h} \) 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-graph (i.e., \( H = \{ v_1, v_2, \ldots, v_{n+3k} \} \)) and the final outputs of the GCN layers (i.e., \( G_L^* = \{ g_1, g_2, \ldots, g_{n+3k} \} \)) into attention mechanism. The process is defined by following transformation,

\[
\begin{align*}
\alpha_i &= \text{softmax}(\sum_{j=1}^{n+3k} v_j^T g_i), \\
f_i &= \sum_{l=1}^{n+3k} \alpha_i h_i,
\end{align*}
\]

where $\alpha_i$ refers to attention scores, $f_i$ 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,

\[
p_l = \text{softmax}(W_l f_i + b_l),
\]

where $p_l \in \mathbb{R}^2$ is the predicted probability vector of LIL, \( W_l \in \mathbb{R}^{d \times 2} \) and \( b_l \in \mathbb{R}^2 \) are trainable parameters. Ultimately, we calculate the cross-entropy loss to supervise our LIL module as follows,

\[
L_{ce} = y^l \log p_l^i + (1 - y^l) \log (1 - p_l^i) + \lambda_1 \| \Theta_l \|_2^2,
\]

where \( y^l \) and \( p_l^i \) are the $i$-th elements of the ground truth $y$ and $p_l$, respectively. \( \lambda_1 \) 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 \( H_m \) defined in Eqn.(5) as original query, the text feature \( H' \) 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,

\[
\begin{align*}
Q &= H_m^T W_l^Q, \\
K &= H' W_l^K, \\
V &= H' W_l^V, \\
H' &= \text{softmax}(\frac{Q^T V}{\sqrt{d_h}}), \\
\hat{H} &= LN(H' + H_m^T), \\
\hat{H} &= LN(H + (W^M H + b^M)),
\end{align*}
\]

where the query $Q \in \mathbb{R}^{(r+1) \times d_h}$ is projected from the encoded visual feature $H_m^T$, the key $K \in \mathbb{R}^{n \times d_h}$ and the value $V \in \mathbb{R}^{n \times d_h}$ are both projected from the encoded textual feature $H'$. $LN(\cdot)$ refers to layer normalization operation (Ba, Kiros, and Hinton 2016), and $H' \in \mathbb{R}^{(r+1) \times d_h}$ is the output of attention mechanism, $H \in \mathbb{R}^{(r+1) \times d_h}$ represents the hidden representations of the residual connection, and $\hat{H} \in \mathbb{R}^{(r+1) \times d_h}$ which includes textual and visual information becomes the new queries to lead the model to extract important information.

In addition, \( \{ W^Q, W^K, W^V, W^M \} \in \mathbb{R}^{d_h \times d_h} \) and $b^M \in \mathbb{R}^{d_h}$ are to-be-learned parameters. We stack $B$ such attention mechanism layers, get the output of the last layer as $\hat{H}_B$ and treat encoding of $[CLS]$ token in the output $\hat{H}_B$ as the final representation $f_g$. Similar to the LIL, we utilize fully connected layers to obtain the predicted probability distributions of GIL $p_g \in \mathbb{R}^2$ as follows,

\[
p_g = \text{softmax}(W_g f_g + b_g),
\]

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

\[
L_{ce}^g = y^g \log P^g + (1 - y^g) \log (1 - P^g) + \lambda_2 \| \Theta_g \|_2^2,
\]

where $y^g$ and $P^g$ are the $i$-th elements of the ground truth $y$ and $p_g$, respectively. $\Theta_g$ denotes all trainable parameters in GIL module and \( \lambda_2 \) 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 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,

\[
\begin{align*}
L_{kl}^{g \rightarrow l} &= \eta_1 D_{KL}(p_g \parallel p_l), \\
L_{kl}^{l \rightarrow g} &= \eta_2 D_{KL}(p_l \parallel p_g),
\end{align*}
\]

where $(g \rightarrow l)$ denotes the knowledge transferring from the GIL to LIL, and similarly, $(l \rightarrow g)$ 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,

\[
\eta_{1/2} = \begin{cases} 
1, & \text{if } \text{argmax}(p_g/l) = Y \\
0, & \text{otherwise}
\end{cases}
\]

where argmax denotes the operation that gains the predicted labels from the predicted result $p_g/l$, and $p_g/l$ is the predicted probability distribution of the module GIL/LIL which shares the knowledge (i.e., $p_g$ for $L_{kl}^{g \rightarrow l}$ and $p_l$ for $L_{kl}^{l \rightarrow g}$). Similar to (Hinton, Vinyals, and Dean 2015), we also sharpen the predicted distribution of the model with a temperature parameter $\tau$ for knowledge transfer.

We combine all loss functions as follows,

\[
\begin{align*}
L' &= L_{ce} + \delta_1 L_{kl}^{g \rightarrow l}, \\
L'' &= L_{ce} + \delta_2 L_{kl}^{l \rightarrow g},
\end{align*}
\]

9511
where $\delta_1$ and $\delta_2$ are non-negative hyper-parameters. Ultimately, the binary classification prediction result $\hat{Y}$ is defined as follows,

$$\hat{Y} = arg\max_{y} (P_y + P_{\neg y} \delta_2).$$

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); InCrossMGs (Liang et al. 2021) and CMGCN (Liang et al. 2022).

Table 1 illustrates the 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.

<table>
<thead>
<tr>
<th>MODALITY</th>
<th>METHOD</th>
<th>Acc (%)</th>
<th>F1-score</th>
<th>Macro-average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pre (%)</td>
<td>Rec (%)</td>
<td>F1 (%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pre (%)</td>
</tr>
<tr>
<td>single-modal</td>
<td>Image (Cai, Cai, and Wan 2019)†</td>
<td>84.44</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>ViT (Dosovitskiy et al. 2021)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>TextCNN (Kim 2014)+*</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>Bi-LSTM†+</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>SIARN (Tay et al. 2018)+*</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>SMSD (Xiong et al. 2019)+*</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>BERT (Devlin et al. 2019)+*</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>RoBERTa (Liu et al. 2019)+*</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td>multi-modal</td>
<td>HFM (Cai, Cai, and Wan 2019)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>D&amp;R Net (Xu, Zeng, and Mao 2020)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>Res-BERT (Pan et al. 2020)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>Att-BERT (Pan et al. 2020)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>InCrossMGs (Liang et al. 2021)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>CMGCN (Liang et al. 2022)†</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>MILNet</td>
<td>84.02</td>
<td>78.51</td>
<td>82.89</td>
</tr>
</tbody>
</table>

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.
We exhibited two testing...in the case of sarcasm detection, we exhibited two testing...

To get an intuitive understanding of how the MILNet works...

On Case Study (RQ3)

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

Table 2: Experiment results of ablation study.

<table>
<thead>
<tr>
<th>MODEL</th>
<th>Acc. (%)</th>
<th>F1 (%)</th>
<th>Macro-F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIL-only</td>
<td>88.00</td>
<td>84.70</td>
<td>87.41</td>
</tr>
<tr>
<td>GIL-only</td>
<td>88.25</td>
<td>85.49</td>
<td>87.81</td>
</tr>
<tr>
<td>w/o-embedded-text</td>
<td>89.29</td>
<td>86.67</td>
<td>88.86</td>
</tr>
<tr>
<td>w/o-mutual-learning</td>
<td>89.67</td>
<td>86.62</td>
<td>89.61</td>
</tr>
<tr>
<td>w/o-sample-screening</td>
<td>88.96</td>
<td>86.76</td>
<td>88.60</td>
</tr>
<tr>
<td>w/o-text-modal-relations</td>
<td>88.05</td>
<td>85.80</td>
<td>87.74</td>
</tr>
<tr>
<td>w/o/image-modal-relations</td>
<td>89.12</td>
<td>86.85</td>
<td>88.79</td>
</tr>
<tr>
<td>w/o-cross-modal-relations</td>
<td>88.96</td>
<td>86.65</td>
<td>88.62</td>
</tr>
<tr>
<td>w/-similarity</td>
<td>89.00</td>
<td>86.81</td>
<td>88.69</td>
</tr>
<tr>
<td>MILNet</td>
<td>89.50</td>
<td>87.11</td>
<td>89.12</td>
</tr>
</tbody>
</table>

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.

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.

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


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


