Multi-Modal Sarcasm Detection via Graph Convolutional Network and Dynamic Network

Jiaqi Hao
College of Computer Science
Inner Mongolia University
Hohhot, China
haojiaqi1277@163.com

Junfeng Zhao*
College of Computer Science
Inner Mongolia University
Hohhot, China
cszjf@imu.edu.cn

Zhigang Wang
College of Computer Science
Inner Mongolia University
Hohhot, China
22309008@mail.imu.edu.cn

Abstract

Sarcasm is a form of language used to convey implicit information contradicting the literal meaning of words, often observed on online social media platforms. Accurately detecting satirical or ironic expressions could significantly enhance sentiment analysis and opinion mining. For multi-modal data, capturing both inter- and intra-modal incongruities is crucial for this task. Recently, graph-based approaches to modeling incongruous features bet-ween image and text have made significant progress in this task. However, these methods rely on static networks to capture incongruous features, which makes them inflexible in adapting to diverse groups of text and image, or neglect important information due to inadequate use of text and image. To address these limitations, we propose a multi-modal sarcasm detection model based on the combination of Graph Convolutional Network and Dynamic Network. The graph convolutional network learns the incongruity of the three modal graphs and makes full use of the object-level information. The dynamic network dynamically captures the incongruity between the global-level image and the text and can flexibly adapt to different image and related text. At the same time, we generate augmented text to better utilize the text information. Extensive experiments demonstrate that our proposed method performs favorably against state-of-the-art approaches.

CCS Concepts

• **Information systems** → **Multimedia information systems**; **Sentiment analysis**; *Clustering and classification*; • **Computing methodologies** → *Information extraction.*

Keywords

Multi-modal Sarcasm Detection, Graph Convolutional Network, Dynamic Network

ACM Reference Format:
Jiaqi Hao, Junfeng Zhao, and Zhigang Wang. 2024. Multi-Modal Sarcasm Detection via Graph Convolutional Network and Dynamic Network. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3627673.3679703

*Corresponding Author.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679703

1 Introduction

Nowadays, internet users often employ various rhetorical strategies to express their thoughts and emotions, such as sarcastic expressions. Sarcasm is a form of language that conveys implicit information or intent, and it has been a longstanding topic in various fields including psychology [25], sociology [32], and neuroscience [16], among others. The literal meaning of sarcasm is typically opposite to its underlying true intent [41]. Accurately detecting sarcasm or sarcastic expressions can help us better understand the emotions and opinions people truly want to convey on social media, and it has wide applications in many areas such as social media analysis, customer service improvement, opinion surveys, sentiment analysis, etc.

Traditional sarcasm detection mainly studies the detection of emotional incongruities from text content [14]. Some early studies concentrated on learning the contextual incongruity with feature engineering approaches [5, 10, 30]. In recent years, neural network-based methods have been widely applied to textual sarcasm detection [2, 6, 38]. There have also been recent studies using external knowledge resources to further capture sentence incongruity [24, 26]. Given the increasing presence of image and text combined with information on social media platforms, traditional sarcasm detection methods are no longer sufficient to meet current needs.

Multi-modal sarcasm detection is designed to recognize the ironic sentiment within multiple modalities [4, 31], and has attracted increasing attention from researchers. Mining inter- and intra-modal incongruities is a key strategy for multi-modal sarcasm detection. In existing studies, some models focus on the use of global-level image features for incongruous learning [18, 22, 28, 33]. Although great progress has been made, there are complex visual details in the image, and simply using the global-level image features is too noisy, which will cause noise interference to the incongruous capture. To address this limitation, some work extracts object-level features as input to the model [19, 35]. While this approach has achieved good performance, but to give up the global-level image features is inappropriate, because it can provide background information about the image content, supplement object-level features. Therefore, abandoning global-level image features can result in information loss, especially for those objects or scenes that are not predefined in the object detection model. Inspired by this, recent work [29] a mutual-enhanced incongruity learning network for

(a) <user> good to see you keep your beans fresh ... ew

(b) just got over the flu now i' m in the hospital fun

(c) thanks for the spoiler <user> ! i didn 't want to find out who won hoh on my own anyway .

Figure 1: Examples of Twitter data with sarcasm. (a) Incongruity between the beans in the image and the text "beans fresh". (b) The text expresses happiness in the hospital, but the factual image contradicts this. (c) The main focus is on textual incongruity; "thanks for the spoiler" contradicts our common knowledge and contains sarcastic implications.

multi-modal sarcasm detection, which aims to fully exploit the object-level and global-level image features. Unfortunately, all the above methods adopt the architecture of static networks, which are not flexible enough to capture different types of incongruities for different image and related text. In real life, there are various types of multi-modal ironic expressions. For example, the object in the image is incongruous with a certain word or phrase in the text, the overall description of the text is incongruous with the facts presented by the image, and the image only serves as a supplement while the text context is incongruous, as shown in Figure 1.

To this end, we propose a multi-modal sarcasm detection model based on the combination of Graph Convolutional Network and Dynamic Network (GCN-DN), which aims to make full use of text features and image features, and also use dynamic network to solve the problem of inflexibility caused by static network. Among them, graph convolutional network can make full use of object-level information and better capture the inter and intra-modal semantic relationships, while dynamic network can dynamically capture the incongruity between the global-level image and the text and can flexibly adapt to different image and related text. It's worth mentioning that to better utilize the information in the text, we simulate human thinking patterns. By using pre-trained commonsense reasoning tools, we supplement the implicit human emotions and potential impacts of events in the text, combining them with the original text content to generate augmented text. This simulation of human thinking enables the method to better understand sarcasm, thereby improving the accuracy of sarcasm detection. The contributions of our work are summarized below.

- For the first time, we propose the use of a combination of graph convolutional networks and dynamic networks for the multi-modal sarcasm detection task.
- We introduce a novel external knowledge enhancement method that simulates human thinking to generate augmented text, to help multi-modal sarcasm detection.
- Experimental results on a public dataset demonstrate the effectiveness of our proposed method for multi-modal sarcasm detection.

2 Related Work

2.1 Multi-Modal Sarcasm Detection

With the development of social media, the detection and understanding of sarcasm needs to consider the relationship between multiple modes. Early studies utilize simple fusion methods of visual and textual information or methods based on decomposition and relation network and attention mechanism for multi-modal sarcasm detection. Schifanella et al. [31] pioneer the use of both textual and visual information, employing a cascaded approach for sarcasm detection. Cai et al. [3] propose a hierarchical fusion model with image features, image attribute features and text features to deal with multi-modal sarcasm detection. Xu et al. [40] construct a decomposition and relation network for multi-modal sarcasm detection. Pan et al. [27] propose inter-modality attention and co-attention to learn the contradiction of sarcasm. In the graphics-based modeling approach, Liang et al. [18] propose to use multi-layer interactive graph convolutional networks to fuse features between text and image and capture multi-modal graph representations. Liang et al. [19] explore a local semantically guided detection approach that explicitly connects important visual areas to text markers. Liu et al. [22] build a hierarchical congruity model based on cross-attention mechanism and graph neural network. Qiao et al. [29] propose a method to model multi-modal sarcasm detection from both local semantic guidance and global perspective. Wei et al. [35] leverage global graph-based semantic awareness to handle this task. In addition, Wen et al. [36] propose a dual inconsistent perception network for multi-modal sarcasm detection, consisting of semantic-enhanced distribution modeling and Siamese Sentiment Contrastive learning modules. Although existing methods have yielded encouraging results, both image and text features are not fully utilized, and the use of static networks limits the flexibility of models.

2.2 Graph Neural Networks

Graph neural network (GNN) is a deep learning model specifically designed to process graph data. It can learn the representation of nodes and edges in the graph and perform prediction and inference of various tasks on this basis. Classical GNN models include Graph Convolution Network (GCN) [15] and Graph Attention Network (GAT) [34]. In recent years, these models have been widely used in the field of multi-modal learning and have achieved good performance in many studies. Such as multi-modal facial expression recognition [39], multi-modal recommendation system [17], multi-modal emotion recognition [43], multi-modal medical image classification [21] and multi-modal sarcasm detection [35].

2.3 Multi-Modal Dynamic Networks

Compared with static networks with fixed calculation graphs and parameters in the inference stage, dynamic networks can adapt their structures or parameters to different inputs, which has significant advantages in accuracy, computational efficiency, and flexibility. Multi-modal dynamic networks also demonstrate strong

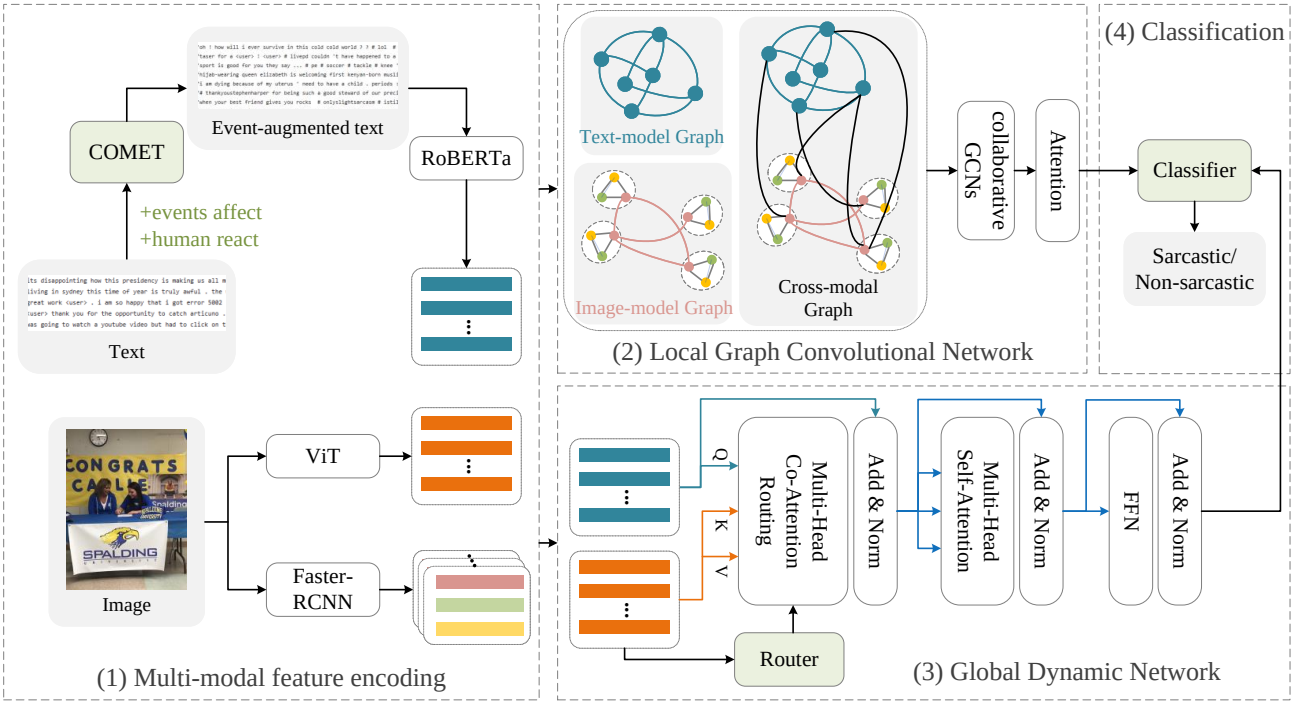


Figure 2: Overall architecture of our proposed GCN-DN for multi-modal sarcasm detection.

performance in multi-modal tasks, such as multi-modal emotion analysis [11, 42], multi-modal Entity Linking [37], and multi-modal 3D target detection [20]. While previous studies have primarily focused on the applications of dynamic networks in social networks, sentiment analysis, and traffic networks, few have explored their potential for multi-modal sarcasm detection tasks.

3 Methodology

This section provides a detailed introduction to our proposed GCN-DN, as shown in Figure 2, which consists of four parts: multi-modal feature encoding, local graph convolutional network, global dynamic network, and classification. We use COMET to generate augmented text and use RoBERTa [23] to encode the text, while using Faster-RCNN [1] and pre-trained Vision Transformer (ViT) [7] to obtain global-level image features and image object-level features. In the local graph convolutional network, we take the extracted text features and object-level features as inputs to construct text-modal graph, image-modal graph, and cross-modal graph, each of which uses the GCN to learn the internal semantic relationship and incongruity. In the global dynamic network, we input the extracted text features and global-level image features into the dynamic network, dynamically learn the interaction relationship between text and image, thereby obtaining a more accurate and rich global incongruity representation. Finally, the results obtained from the local graph convolutional network and the global dynamic network are fed into the classifier for final multi-modal sarcasm detection.

3.1 Multi-Modal Feature Encoding

Our work mainly studies multi-modal sarcasm detection with text and image input, that is, for a given N training samples $D = \{s^1, s^2, \ldots, s^N\}$, each sample s^i has two inputs: $\mathbf{Text}^i$ and $\mathbf{Image}^i$.

3.1.1 Text Encoding. COMET [12] is a pre-trained commonsense inference tool that infers various commonsense relationships associated with relevant events of a given text. Through referring to the prior psychological, cognitive, and linguistic literature [8, 9, 13], clearly, sarcasm is always associated with the impact of social events and human emotions [26]. Humans quickly determine sarcasm because our brains possess background knowledge and common sense about sarcastic scenarios, allowing us to directly comprehend the entire sarcastic text and thereby capture the true emotions hidden in the given text. Therefore, to fully utilize the latent information of the text in the multi-modal samples, for the original text $\mathbf{Text}^i = \{w_1^i, w_2^i, \ldots, w_M^i\}$, where w_j^i represents the j-th token in the text and M is the total number of tokens, we input it into COMET, simulate human thinking, and obtain two sequences of implicit social event impacts and human emotions, represented as $\mathbf{effect}^i = \{\tilde{w}_1^i, \tilde{w}_2^i, \ldots, \tilde{w}_{\tilde{M}}^i\}$ and $\mathbf{react}^i = \{\overline{w}_1^i, \overline{w}_2^i, \ldots, \overline{w}_{\overline{M}}^i\}$. The original text $\mathbf{Text}^i$ and the two obtained sequences are concatenated to obtain the augmented text, denoted as $\mathbf{A\_Text}^i = \mathbf{Text}^i \oplus \mathbf{effect}^i \oplus \mathbf{react}^i$, where $\oplus$ denotes the concatenation operator. We encode the resulting augmented text into RoBERTa to obtain the text feature $\mathbf{T}$.

$$\mathbf{T} = [t_1, t_2, \ldots, t_m] = \text{RoBERTa}(A_{\text{Text}}), \tag{1}$$

where $t_i \in \mathbb{R}^{d_t}$ represents the hidden state vector of the i-th token in the text, d_t is the dimension of the hidden representation, and $m = M + \tilde{M} + \bar{M}$ is the total number of tokens in the augmented text.

3.1.2 Image Encoding. To achieve more comprehensive visual information, we extract both object-level and global-level visual features.

For object-level feature extraction, we rely on Faster R-CNN to detect objects and extract their features. To ensure the quality of extracted object-level features, we select only the top k objects with the highest confidence score for feature extraction. Each object obtains a visual feature $\mathbf{v}_i \in \mathbb{R}^{d_v}$, a positional feature $\mathbf{p}_i \in \mathbb{R}^{d_p}$, an object class c_i and an object attribute a_i. $i \in [1, k]$ represents the i-th object, d_v and d_p denote the dimension of visual and positional features, respectively. The visual feature $\mathbf{v}_i$ and positional feature $\mathbf{p}_i$ are fused through a linear projection to obtain a richer visual feature, denoted as $\mathbf{f}_i = \mathbf{W}_v\mathbf{v}_i + \mathbf{W}_p\mathbf{p}_i + \mathbf{b}_f$, where $\mathbf{W}_v \in \mathbb{R}^{d_f \times d_v}$ and $\mathbf{W}_p \in \mathbb{R}^{d_f \times d_p}$ are learnable weight matrices, and $\mathbf{b}_f \in \mathbb{R}^{d_f}$ is a bias term. Based on Eq. (1), we utilize RoBERTa to transform the object class c_i and object attribute a_i into their vector representations $\tilde{\mathbf{c}}_i$ and $\tilde{\mathbf{a}}_i$, respectively. Combining $\mathbf{f}_i$, $\tilde{\mathbf{c}}_i$ and $\tilde{\mathbf{a}}_i$ together yields the feature representation of the i-th object. The object-level representation of the entire image is represented by a matrix $\mathbf{V}_o \in \mathbb{R}^{3k \times d_f}$ as follows,

$$\mathbf{V}_o = [[\mathbf{f}_1, \tilde{\mathbf{c}}_1, \tilde{\mathbf{a}}_1]^\top, [\mathbf{f}_2, \tilde{\mathbf{c}}_2, \tilde{\mathbf{a}}_2]^\top, \ldots, [\mathbf{f}_k, \tilde{\mathbf{c}}_k, \tilde{\mathbf{a}}_k]^\top], \tag{2}$$

Where each row corresponds to the feature representation of an object in the image, and the feature of the i-th object can be expressed as $\mathbf{I}_i = [\mathbf{f}_i, \tilde{\mathbf{c}}_i, \tilde{\mathbf{a}}_i]^\top$.

To obtain global-level image features, firstly, we resize the $\mathbf{Image}^i \in \mathbb{R}^{l \times w}$ and divide it into r flat patches, such that the entire image can be represented as a sequence containing r patches. Next, this sequence is sent to ViT as input for processing, and the global-level image features $\mathbf{V}$ is obtained, expressed as,

$$\mathbf{V} = [v_1, v_2, \ldots, v_r] = \text{ViT}(\mathbf{Image}), \tag{3}$$

where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ is the image embedding of the i-th patch in the image and d_v is the dimension of the image embedding.

3.2 Local Graph Convolutional Network (LGCN)

Compared to single-modal data, multi-modal data offers richer semantic relationships. Ironic expressions can appear solely in either text or image, or from contrasting descriptions across multiple modalities. To capture these inter- and intra-modal incongruities, we construct text, image, and cross-modal graphs, using GCN to better learn and identify ironic expressions.

3.2.1 Text-modal graph. To extract ironic expressions from the textual perspective and capture semantic associations between words in sentences, we construct a text modal graph $\mathcal{G}^t$. In this graph, each token corresponds to a node, represented as $\mathbf{v}_i^t$, where $i \in [1, m]$ denotes the i-th token in the augmented text. The node features are initialized using the $\mathbf{t}_i$ obtained from Eq. (1). The edges of graph are defined by the dependency tree[1], an efficient method as demonstrated in previous studies [18]. Concretely, based on the parent node of each token in the dependency tree, we connect the corresponding nodes in the text-modal graph to represent token dependencies. The text-modal adjacency matrix $\mathbf{A}^t \in \mathbb{R}^{m \times m}$ is constructed as follows,

$$\mathbf{A}^t_{i,j} = \begin{cases} 1, & \text{if } D(\mathbf{t}_i, \mathbf{t}_j) = 1, \quad i, j \in [1, m] \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where $D(\mathbf{t}_i, \mathbf{t}_j)$ indicates that token $\mathbf{t}_i$ and token $\mathbf{t}_j$ have a certain dependency relationship in the dependency tree, the corresponding nodes $\mathbf{v}_i^t$ and $\mathbf{v}_j^t$ are connected. To enrich the dependency information of the text, we construct the graph as an undirected graph (i.e., $\mathbf{A}^t_{i,j} = \mathbf{A}^t_{j,i}$) and set a self-loop for each token (i.e., $\mathbf{A}^t_{i,i} = 1$).

3.2.2 Image-modal graph. There are also some visual features in ironic expressions that are separate from text. To capture the visual semantic association between different objects in the image, we construct an image-modal graph $\mathcal{G}^v$. In the graph, there are a total of k objects, each with three nodes represented as $\mathbf{v}_i^v$, where $i \in [1, 3k]$, corresponding to the three feature vectors of the object: richer visual features, object class, and object attributes. The node set can be initialized as $\{\mathbf{f}_1, \tilde{\mathbf{c}}_1, \tilde{\mathbf{a}}_1, \ldots, \mathbf{f}_k, \tilde{\mathbf{c}}_k, \tilde{\mathbf{a}}_k\}$, where the object class serves as the representative node connecting the objects. As the three nodes of each object essentially describe the same object from different aspects, we fully connect them. The construction of edges between objects is based on the Intersection over Union (IoU) scores between objects to reflect their spatial relationship and semantic correlation. The image-modal adjacency matrix $\mathbf{A}^v \in \mathbb{R}^{3k \times 3k}$ is constructed as follows,

$$\mathbf{A}^v_{i,j} = \begin{cases} 1, & \text{if } i \bmod 3 = j \bmod 3, \quad i, j \in [1, 3k] \\ S_{i,j}, & \text{if } i \bmod 3 = 1, j \bmod 3 = 1, \quad i, j \in [1, 3k] \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where *mod* denotes the modulo operation, which links the three nodes corresponding to the same object. *rem* denotes the remainder operation, used for linking different objects. $S_{i,j}$ is the IoU score, serving as the weight of the edges between different objects.

3.2.3 Cross-modal graph. Many times, irony is expressed by combining image and text. To integrate the semantic relationship between image and text for a more comprehensive sarcasm detection, we construct a cross-modal graph $\mathbf{G}^c$. Its nodes cover all text tokens and object-level image features, denoted as v_i^c, where $i \in [1, m+3k]$. The node set is initialized as $\{t_1, t_2, \ldots, t_m, f_1, \tilde{c}_1, \tilde{a}_1, \ldots, f_k, \tilde{c}_k, \tilde{a}_k\}$. To construct edges in the cross-modal graph, the process involves two steps: 1) We use the association information in knowledge graph ConceptNet5[2] to judge whether there is a correlation between text and image. If there is a correlation, the nodes are connected; 2) We integrate the existing edges in the text-modal graph and the image-modal graph to enrich the information in the cross-modal graph. This means that if edges representing semantic relationships between words or objects already exist in text-modal graph and image-modal graph, those edges will also be included in cross-modal graph. The cross-modal adjacency matrix $\mathbf{A}^c \in \mathbb{R}^{(m+3k)\times(m+3k)}$ is

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

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

constructed as follows,

$$\mathbf{A}^c_{i,j} = \begin{cases} \mathbf{A}^t_{i,j}, & \text{if } \mathbf{A}^t_{i,j} > 0,\ i,j \in [1,m] \\ \mathbf{A}^v_{i,j}, & \text{if } \mathbf{A}^v_{i,j} > 0,\ i,j \in [m+1, m+3k] \\ 1, & \text{if } K(t_i, I_j),\ i \in [1,m],\ j \in [m+1, m+k] \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where, $K(t_i, I_j)$ denotes that there exists a certain relationship between the token t_i and objects I_j in ConceptNet5.

3.2.4 Graph Convolutional Network. We employ a multi-layer collaborative GCNs architecture to learn inter- and intra-modal incongruous expressions. For each collaborative GCNs layer, we construct text-modal layer, image-modal layer, and cross-modal layer, facilitating collaborative learning across the three modal graphs. These modal graphs are employed to adjust and refine the graphical representations in multi-modal sarcasm detection. The specific process is defined as follows,

$$\begin{cases} \mathbf{G}^t_l = \text{ReLU}(\tilde{\mathbf{A}}^t \mathbf{G}^c_{l-1} \mathbf{W}^t_l + \mathbf{b}^t_l) \\ \mathbf{G}^v_l = \text{ReLU}(\tilde{\mathbf{A}}^v \mathbf{G}^t_l \mathbf{W}^v_l + \mathbf{b}^v_l) \\ \mathbf{G}^c_l = \text{ReLU}(\tilde{\mathbf{A}}^c \mathbf{G}^v_l \mathbf{W}^c_l + \mathbf{b}^c_l), \end{cases} \tag{7}$$

where G^x_l are the representations of nodes in corresponding graphs after the l-th collaborative GCNs process. $\mathbf{W}^x_l \in \mathbb{R}^{d_h \times d_h}$ and $\mathbf{b}^x_l \in \mathbb{R}^{d_h}$ are the trainable parameters of the l-th collaborative GCNs layer, where $x \in \{t, v, c\}$ and $l \in [1, L]$. $\tilde{\mathbf{A}}^x = (\mathbf{D}^x)^{-\frac{1}{2}} \mathbf{A}^x (\mathbf{D}^x)^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, where $\mathbf{D}^x$ is the degree matrix of $\mathbf{A}^x$. The original input nodes of the first interactive GCNs layer are the combination of text representation and object-level image representation, i.e. $\mathbf{G}^c_1 = \{t_1, t_2, \ldots, t_m, f_1, \tilde{c}_1, \tilde{a}_1, \ldots, f_k, \tilde{c}_k, \tilde{a}_k\}$.

Subsequently, inspired by [18, 19], we employ a retrieval-based attention mechanism to capture attention features from both text and image. We input the initial node representations of cross-modal graph (i.e., $\mathbf{H} = \{\mathbf{v}^c_1, \mathbf{v}^c_2, \ldots, \mathbf{v}^c_{m+3k}\}$) and the final outputs of the collaborative GCNs layers (i.e., $\mathbf{G}^c_L = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{m+3k}\}$) into attention mechanism. For each node $\mathbf{v}_j$, we calculate its attention scores with respect to other nodes, which are then used to assess the importance of nodes in the graph. Using these attention scores, we perform a weighted summation of node representations to obtain the final sarcasm representation of LGCN for sarcasm detection, as follows,

$$f_G = \sum_{i=1}^{m+3k} \alpha_i h_i, \tag{8}$$

where $\alpha_i = \text{softmax}\left(\sum_{j=1}^{m+3k} \mathbf{v}^T_j \mathbf{g}_j\right)$ refers to attention scores, $\mathbf{h}_i$ denotes the representation of the initial node $\mathbf{v}_j$ in the cross-modal graph. Then we feed $\mathbf{f}_G$ into fully connected layers to gain the predicted probability distributions as follows,

$$\mathbf{p}_G = \text{softmax}(\mathbf{W}_G \mathbf{f}_G + \mathbf{b}_G), \tag{9}$$

where $\mathbf{p}_G \in \mathbb{R}^2$ represents the predicted probability vector of LGCN, $\mathbf{W}_G \in \mathbb{R}^{d \times 2}$, $\mathbf{b}_G \in \mathbb{R}^2$ are trainable parameters. Ultimately, we calculate the cross-entropy loss to measure the difference between the predicted probability vector and the ground truth as follows,

$$L^G_{\text{ce}} = y^i \log p^i_G + (1 - y^i) \log(1 - p^i_G) + \lambda_G \|\Theta_G\|^2, \tag{10}$$

where y^i is the i-th element of the ground truth $\mathbf{y}$. Θ_G denotes all trainable parameters in the LGCN module, and λ_G represents the weight coefficients of the Frobenius norm.

3.3 Global Dynamic Network (GDN)

For multi-modal sarcasm detection, capturing cross-modal incongruity is crucial. Previous methods relied on static network architectures, which lack flexibility. Moreover, object-level features and global-level image features have been demonstrated as the optimal choices for fully leveraging image information. Nevertheless, prior research predominantly concentrated on utilizing either one of them exclusively. Therefore, we employ a dynamic network to address this limitation. In the previous LGCN section, we utilized object-level features. In this section, we use global-level image features as inputs to the GDN.

The GDN is a multi-layered network structure that takes augmented text and global-level image as input, and processing and transforming these features through a series of dynamic layers. These dynamic layers perform hierarchical co-attention between text and image, conditioned on different inputs, thus progressively refining and optimizing the representation of multi-modal data. Each dynamic layer mainly consists of three modules: multi-head co-attention routing (MHCAR) module, multi-head self-attention (MHA) module, and feed-forward network (FFN).

3.3.1 Multi-Head Co-attention Routing. The MHCAR module in the dynamic layer employs a parallel multi-head attention mechanism to effectively capture the complex relationships between text and image. By concurrently executing multiple attention heads and weighting their results based on routing probability weights, the MHCAR module achieves fine-grained attention and integration across different components.

Specifically, first, the output of the previous dynamic layer $\mathbf{D}_{l-1}$ and the global-level image features $\mathbf{V}$ are separately subjected to linear transformations to derive the query $\mathbf{Q}_{i,j,l}$, key $\mathbf{K}_{i,j,l}$, and value $\mathbf{V}_{i,j,l}$ for the i-th attention head, as follows,

$$\begin{cases} \mathbf{Q}_{i,j,l} = \mathbf{D}_{l-1} \mathbf{W}^{\mathbf{Q}}_{i,j,l} \\ \mathbf{K}_{i,j,l} = \mathbf{V} \mathbf{W}^{\mathbf{K}}_{i,j,l} \\ \mathbf{V}^l_{i,j} = \mathbf{V} \mathbf{W}^{\mathbf{V}}_{i,j,l}, \end{cases} \tag{11}$$

where $\mathbf{W}^Q_{i,j,l} \in \mathbb{R}^{d_t \times d_h}$, $\mathbf{W}^K_{i,j,l} \in \mathbb{R}^{d_v \times d_h}$ and $\mathbf{W}^V_{i,j,l} \in \mathbb{R}^{d_v \times d_h}$ are parameter matrices. Specifically, the initial input $\mathbf{D}_0$ is the textual input $\mathbf{T}$ obtained from Eq. (1). Next, we compute the attention weights and attention distribution between text and image, and then obtain the representation of the i-th attention head head^l_i, in the l-th dynamic layer. The calculation is as follows,

$$\begin{cases} \mathbf{head}^l_i = \sum_{j=0}^{p_l - 1} \alpha^l_j \, \mathbf{CA}^l_{i,j}(\mathbf{Q}_{i,j,l}, \mathbf{K}_{i,j,l}, \mathbf{V}^l_{i,j}, \mathbf{A}^j) \\ \mathbf{CA}^l_{i,j} = \text{softmax}\left(\dfrac{\mathbf{Q}_{i,j,l} \mathbf{K}^\top_{i,j,l}}{\sqrt{d_h}} \otimes \mathbf{A}^j\right) \mathbf{V}^l_i, \end{cases} \tag{12}$$

where $\text{CA}^l_{i,j}$ is the co-attention function used to compute the attention distribution between the text and image, indicating the degree

of attention between the two modalities, and p_l represents the number of co-attention functions in the l-th layer. $\mathbf{Q}_{i,l}^{\top}\mathbf{K}_{i,l} \in \mathbb{R}^{m\times r}$ represent the attention matrices between the two modalities, and $\otimes$ denotes element-wise matrix product. $\mathbf{A}^j = [\mathbf{v}_1^s, \mathbf{v}_1^s, \ldots, \mathbf{v}_1^s] \in \mathbb{R}^{m\times r}$ denotes the co-attention mask matrix, composed of masking vectors $\mathbf{v}_l^s \in \mathbb{R}^r$ (where, $l \in [1, r]$) generated by applying a sliding window method of size $(2s+1)\times(2s+1)$ to restrict the region the text can see. We gradually increase the capture of incongruities between the text and image by constructing different co-attention mask matrices as the dynamic layers increase. α_j^l is the routing probability weight, determining the influence of the j-th co-attention function in the entire attention mechanism. The process is defined by following transformation,

$$\boldsymbol{\alpha}_j^l = \text{gumbel softmax}\left(\text{MLP}\left(\text{APool}(\mathbf{I})\right)\right) \in \mathbb{R}^{p_l}, \tag{13}$$

where APool is an adaptive average pooling operation and MLP is a two-layer multilayer perceptron with hidden dimension d_r. Finally, based on the head_i^l obtained from Eq. (12), the final output of MHCAR can be calculated as follows,

$$\mathbf{MHCAR}_l(\mathbf{D}_{l-1}, \mathbf{V}) = \text{concat}_{i=1}^{h}\left(\mathbf{head}_i^l, \mathbf{O}_T^l\right), \tag{14}$$

where $\text{concat}_{i=1}^{h}$ denotes the concatenation operation performed on the outputs of h attention heads. $\mathbf{O}_T^l \in \mathbb{R}^{d_o\times d_o}$ serves as the projection matrix, aiming to map the concatenated outputs of attention heads head_i^l to a new space, ensuring compatibility with the subsequent layer. A residual connection and a normalization layer (LN) follow the MHCAR module, resulting in the output of the l-th dynamic layer as follows,

$$\mathbf{D}_l^{\text{MHCAR}} = \text{LN}\left(\mathbf{MHCAR}_l(\mathbf{D}_{l-1}, \mathbf{V}) + \mathbf{D}_{l-1}\right), \tag{15}$$

where $\mathbf{D}_{l-1}$ is the output of the $l-1$ dynamic layer.

3.3.2 Multi-Head Self-Attention. To enhance the network hierarchy, improve feature representation capabilities, and better extract associative information between text and image, thereby improving the expression power and generalization ability of the model, we introduce the MHA module.

Specifically, in the l-th dynamic layer, the MHA module takes the output $\mathbf{D}_l^{\text{MHCAR}}$ from the MHCAR module as input. It then performs linear transformations to derive the query $\mathbf{Q}_{i,j,l}$, key $\mathbf{K}_{i,j,l}$, and value $\mathbf{V}_{i,j}^l$ for the i-th attention head, following a process similar to that described in Eq. (11). For each head, the attention score is calculated, and the representation of the i-th attention head in the l-th dynamic layer is as follows,

$$\mathbf{head}_i^l = \text{softmax}\left(\frac{\mathbf{Q}_{i,l}\mathbf{K}_{i,l}^{\top}}{\sqrt{d_h}}\right). \tag{16}$$

The outputs of each head_i^l are combined using an attention-weighted sum. Following the MHA module, a residual connection and a LN are applied. The output of the MHA module in the l-th dynamic layer is obtained as follows,

$$\mathbf{D}_l^{\text{MHA}} = \text{LN}\left(\mathbf{MHA}_l\left(\mathbf{D}_l^{\text{MHCAR}}\right) + \mathbf{D}_l^{\text{MHCAR}}\right). \tag{17}$$

Table 1: Statistics of the experimental data.

	Training	Development	Testing
Positive	8642	959	959
Negative	11174	1451	1450
Total	19816	2410	2409

3.3.3 Feed-Forward Network. he FFN module is used in the dynamic layer to process the output of the MHA module by performing two linear transformations, which can be formulated as follows,

$$\mathbf{FFN}_l(\mathbf{D}_l^{\text{MHA}}) = \text{ReLu}(\mathbf{D}_l^{\text{MHA}}\mathbf{W}_1^l + \mathbf{b}_1^l)\mathbf{W}_2^l + \mathbf{b}_2^l, \tag{18}$$

where $\mathbf{D}_l^{\text{MHA}}$ is the output of MHA module in the l-th dynamic layer, $\mathbf{W}_1^l \in \mathbb{R}^{d_t\times d_h}$ and $\mathbf{W}_2^l \in \mathbb{R}^{d_h\times d_t}$ are the weight matrices, and $\mathbf{b}_1^l \in \mathbb{R}^{d_h}$ and $\mathbf{b}_2^l \in \mathbb{R}^{d_t}$ are bias terms. After the FFN module, a residual connection and a LN are applied. The output of the FFN module in the l-th dynamic layer also serves as the final output of the l-th dynamic layer, as follows,

$$\mathbf{D}_l = \mathbf{D}_l^{\text{FFN}} = \text{LN}\left(\mathbf{FFN}_l\left(\mathbf{D}_l^{\text{MHA}}\right) + \mathbf{D}_l^{\text{MHA}}\right). \tag{19}$$

Each dynamic layer completes one iteration after processing through the MHCAR, MHA and FFN modules. Upon completing L iterations, the final representation $\mathbf{f}_D$ (where $\mathbf{f}_D = \mathbf{D}_L$) for sarcasm detection is obtained. Similar to the LGCN, a fully connected layer is employed to derive the prediction probability distribution $\mathbf{p}_D \in \mathbb{R}^2$. The cross-entropy loss function L_{ce}^D for the GDN is then defined, having a form analogous to Eq. (12) (13).

3.4 Classification

To fuse the prediction results of the models, we employ a weighted average of the prediction outputs from the LGCN and the GDN. This approach allows us to comprehensively consider the predictive capabilities of both models, thereby achieving improved prediction performance through judicious allocation of weights. The final prediction result $\hat{\mathbf{Y}}$ and the loss function L_f are defined as follows,

$$\hat{\mathbf{Y}} = \text{argmax}\left(\alpha \times \mathbf{p}_G + (1-\alpha)\times \mathbf{p}_D\right), \tag{20}$$

$$L_f = \beta \times L_{\text{ce}}^G + (1-\beta)\times L_{\text{ce}}^D, \tag{21}$$

where, α and β are weights controlling the predictions and loss functions of the LGCN and the GDN, respectively. $\mathbf{p}_G$ and $\mathbf{p}_D$ represent the prediction probability distributions of the LGCN and the GDN, while L_{ce}^G and L_{ce}^D denote the cross-entropy losses of the LGCN and the GDN, respectively.

4 Experiments

4.1 Datasets

We evaluate our approach on the publicly available multimodal sarcasm detection benchmark dataset [3]. Each example in this dataset comprises a text component and a corresponding image. Samples labeled with hashtags such as "#sarcasm" are categorized positive examples, while those without such labels are deemed negative examples. The dataset is divided into a training set, a development set, and a testing set with a ratio of 80%:10%:10%. The statistics of the dataset are shown in Table 1.

Table 2: Performance comparison among different methods on the multi-modal sarcasm dataset in terms of Acc, F1-score, and Macro-average F1-score. The best results are represented in bold. The second-best results are underlined.

Modality	Method	Acc (%)	Pre (%)	Rec (%)	F1 (%)	Macro-average		
						Pre (%)	Rec (%)	F1 (%)
single-model	ResNet	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	73.72	65.56	71.64	68.46	72.78	73.37	72.97
	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
	RoBERTa	85.51	78.24	88.11	82.88	84.83	85.95	85.16
multi-model	HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	Res-BERT	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	HKEmodel	87.36	81.84	86.48	84.09	-	-	-
	CMGCN	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	MILNet	89.50	85.16	89.16	87.11	88.88	89.44	89.12
	DIP	89.59	87.76	86.58	87.17	88.46	89.13	89.01
	G ^{2} SAM	<u>90.48</u>	<u>87.95</u>	<u>89.02</u>	<u>88.48</u>	<u>89.44</u>	<u>89.79</u>	<u>89.65</u>
	DN-GCN	**91.53**	**90.44**	**90.79**	**89.20**	**89.58**	**89.80**	**90.68**

4.2 Experimental Settings

To ensure fairness, we follow previous works [3] for dataset preprocessing. We utilized comet-atomic-2020[3] and roberta-base[4] for augmented text generation, where each word embedding is of dimensionality 768. The image size is adjusted to a resolution of 224×224, and visual embeddings are generated using vit_base_patch32_224[5]. We employ bottom-up-attention[6] for object detection, with a maximum of 10 visual regions. In the LGCN, the number of collaborative GCNs layers is 2, while in the GDN, there are 4 dynamic layers. We optimized our model using Adam. In the LGCN, the learning rate is 10^{-4}, and the weight decay is 10^{-4}. In the GDN, the learning rate is 10^{-6}, and the weight decay is 0.01. We evaluate the model performance using Accuracy, Precision, Recall, and F1 score. Our experimental results are averaged over ten runs with different random seeds.

4.3 Comparison Models

To validate the performance of GCN-DN, we compare it with representative methods of existing single-modal and multi-modal baselines.

4.3.1 Single-Modal Baselines. Image-modality methods utilize visual information for sarcasm detection, including image embeddings with pooling layers using ResNet [3] and pre-trained visual model ViT [7] based on the Transformer architecture. Text-modality methods rely on textual information for sarcasm detection, including text encoding with Bi-LSTM [29] and models based on pre-trained Transformer architecture, such as BERT [6] and RoBERTa [23].

[3]https://github.com/allenai/comet-atomic-2020.
[4]https://huggingface.co/roberta-base.
[5]https://github.com/rwightman/pytorch-image-models.
[6]https://github.com/peteanderson80/bottom-up-attention.

4.3.2 Multi-Modal Baselines. Multi-modal methods that combine image and text information to detect sarcasm, we consider comparing them with the following methods. HFM [3] proposes a hierarchical fusion model, which integrated text, image and image attribute information. Res-BERT [27] connects image features and BERT-based text features to detect sarcasm. Att-BERT [27] proposes different attention strategies to detect sarcasm. InCrossMGs [18] explores an interactive graph convolution network structure to learn the incongruity relations of in-modal and cross-modal graphs jointly and interactively. CMGCN [19] proposes cross-modal graphs based on attribute-object pairs of image objects to capture sarcastic clues. HKEmodel [22] uses image captions as external knowledge to enhance the ability of multi-modal sarcasm detection. MILNet [29] designs a local semantic-guided incongruity learning module and a global incongruity learning module to mutually enhance the ability of multi-modal sarcasm detection. DIP [36] introduces a dual inconsistency perception network consisting of two branches to explore sarcasm information from both factual and emotional aspects. G^{2}SAM [35] proposes a multi-modal sarcasm detection inference paradigm based on graph-based global semantic perception.

4.4 Main Results

We compared the experimental results of different models in the multi-modal sarcasm detection task, as shown in Table 2. From these results, we can draw the following conclusions. 1) The base-line model based on the text modality outperforms the baseline model based on the image modality. This may be because higher information density in text, which provides more explicit sarcasm cues compared to image information. This conclusion suggests that our proposed method of augmented text mining for more text modality information is reasonable and effective to a certain extent. 2) It is evident that multi-modal methods outperform single-modal methods in sarcasm detection tasks. This is because multi-modal methods

Table 3: Experiment results of ablation study.

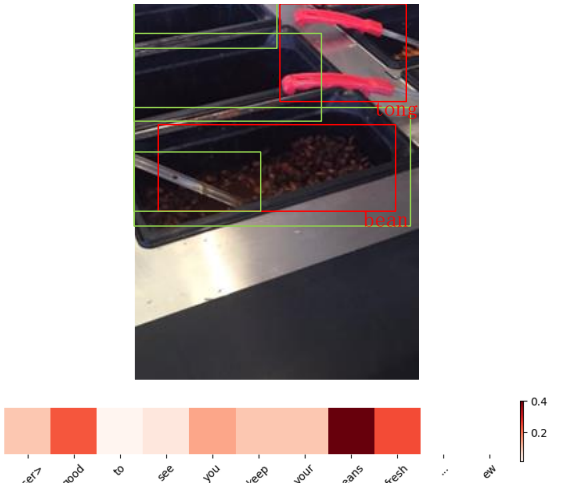
Model	Acc (%)	F1 (%)	Macro-F1 (%)
LGCN-only	88.41	85.11	87.59
GDN-only	88.33	85.46	87.54
- GDN, +Standard Transformer	88.09	84.21	87.36
GDN (p_l=L-1)	91.06	88.70	90.00
GDN (p_l=L-2)	90.73	88.44	89.76
w/o-object-level-features	87.91	84.38	86.81
w/o-global-level-features	87.69	84.24	86.78
w/o-augmented-text	90.95	87.37	90.52
GCN-DN	**91.53**	**89.20**	**90.68**

can effectively leverage both text and image information, thereby enhancing the detection sarcasm. This indicates that capturing both inter- and intra-modal incongruities is crucial for extracting sarcasm clues from both image and text. 3) Our model achieved the best performance across all metrics, outperforming several graph-based baseline models [18, 19, 29, 35]. This suggests that exploring the framework based on graph convolutional networks and dynamic networks holds great potential. We conducted a significance test between GCN-DN and the baseline model G^{2}SAM, and the results showed that GCN-DN significantly outperformed G^{2}SAM on most evaluation metrics (with p-value < 0.05). This validates the superiority of GCN-DN over existing methods.

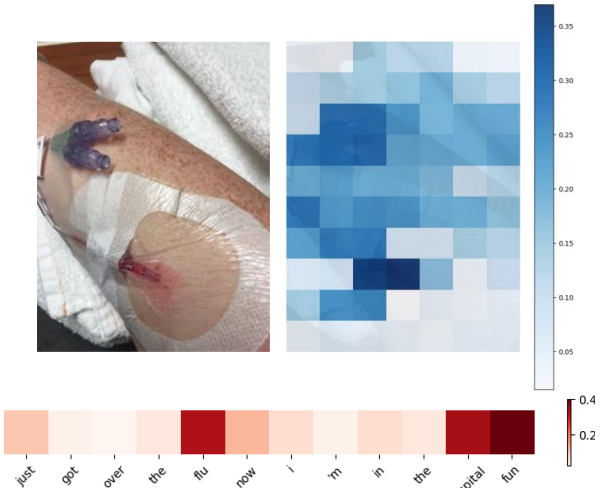
4.5 Ablation Study

This validates the superiority of GCN-DN over existing methods. To analyze the impact of different components of our proposed GCN-DN, we conduct an ablation study. 1) To explore the effect of combining graph convolutional network and dynamic network, we remove the LGCH and the GDN, respectively. 2) To demonstrate the benefits of using the dynamic network, we replaced it with a standard multi-modal transformer. 3) Further-more, we vary the number of co-attention mask matrices in the GDN to analyze its effectiveness. 4) To verify whether using both object-level features and global-level features simultaneously leads to better results, we modified the GCN-DN to exclusively utilize either object-level features or global-level features. 5) To validate the need for augmented text, we discarded it and only fed the original sentence from text modality into our GCN-DN framework.

Table 3 reports the results of the ablation experiment. From the data provided, we have the following observations. 1) Our model surpasses using either the graph convolutional network or the dynamic network alone, indicating that combining the graph convolutional network with the dynamic network can leverage their respective strengths and complement the weaknesses of each, thereby enhancing the model performance. 2) After replacing the dynamic network with the standard multi-modal transformer, we find that removing the dynamic ability of the model leads to performance degradation, which reflects the advance of our proposed dynamic network in capturing cross-modal incongruity. 3) We explored the impact of the number of co-attention mask matrices in the GDN on model performance. We found that increasing the number of co-attention mask matrix types with the growth of dynamic layers



(a) <user> good to see you keep your beans fresh ... ew



(b) just got over the flu now i' m in the hospital fun

Figure 3: Example of case study, (a) is the sample that performs well in LGCN module, and (b) is the sample that performs well in GDN module.

improves performance, while reducing the types of co-attention mask matrices leads to a decline in performance. This indicates that increasing the diversity of co-attention mask matrix types gradually enhances flexibility and generalization ability, helping to better capture cross-modal incongruity between image and text. This further illustrates the advantage of dynamic networks in multi-modal sarcasm detection tasks. 4) We modified our model to use only global-level image features, discarding object-level features in the construction of the graph in the LGCN, similar to the approach of the InCrossMGs model. However, we found that the model performance degraded. Furthermore, when we changed the model to use only object-level features, utilizing object-level features as inputs in the GDN, a decrease in performance is observed. Experimental

results indicate that using both object-level features and global-level image features can enhance model performance. These two types of image features can effectively leverage image information, which is crucial for capturing cross-modal incongruity. 5) After the augmented text containing the effects of events and human emotions was removed, the model performance decreased. This indicates that the generated augmented text, which mimics human thought processes, mines more information from the text modality, aiding the model in better understanding the emotions and contexts within the text.

4.6 Case Study

This confirms the superiority of GCN-DN over existing methods. The key to multi-modal sarcasm detection is to capture the incongruous information between different modalities. Therefore, we present attention visualizations for two test samples, representing distinct types of sarcasm, both necessitating simultaneous consideration of textual and visual cues for effective detection. The findings, depicted in Figure 3, reveal that instances where an object in the image contradicts a specific word or phrase in the text (a), or where the overall textual description is incongruous with the depicted facts in the image (b), are more likely to be attended to by our model. Specifically, in Figure 3(a), the LGCN emphasizes the object region of beans in the image, alongside textual mentions of "beans" and "fresh". In Figure 3(b), despite the absence of specific objects in the image relevant to the textual content, the GDN directs attention towards the area associated with the injection wound, with higher attention scores assigned to "flu" and "fun" in the text.

5 Conclusion

In this paper, we propose a model named GCN-DN, which combines Graph Convolutional Network and Dynamic Network, to capture both inter- and intra-modal incongruities in image and text for multi-modal sarcasm detection tasks. The GCN-DN model, inspired by human cognitive processes, generates augmented text to thoroughly mine information from the textual modality. The integration of graph convolutional network and dynamic network fully utilizes both object-level and global-level image features. Moreover, the design of the dynamic network allows for the capture of sarcasm cues based on varying image and text inputs. Experimental results on public datasets demonstrate the effectiveness of our approach, and ablation studies confirm the superiority and significance of combining graph convolutional network and dynamic network. The equivalence of our proposed method for generating augmented text simulating human thought processes to human cognitive processes requires broader validation and evidence. In the future, we will explore more techniques to integrate different modalities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61962039), the Project on Collaborative Computing Service Failure Prediction and Recovery between Cloud and Edge (Grant No.2023ZD18), and the Fund of Supporting the Reform and Development of Local Universities (Disciplinary Construction).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th international conference on computational linguistics*. 225–243.

[3] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2506–2515.

[4] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4619–4629.

[5] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the fourteenth conference on computational natural language learning*. 107–116.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[8] Raymond W Gibbs Jr. 2002. A new look at literal meaning in understanding what is said and implicated. *Journal of pragmatics* 34, 4 (2002), 457–486.

[9] Rachel Giora and Ofer Fein. 1999. Irony: Context and salience. *Metaphor and symbol* 14, 4 (1999), 241–257.

[10] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 581–586.

[11] Jing He, Haonan Yanga, Changfan Zhang, Hongrun Chen, and Yifu Xua. 2022. Dynamic Invariant-Specific Representation Fusion Network for Multimodal Sentiment Analysis. *Computational Intelligence and Neuroscience* 2022, 1 (2022), 2105593.

[12] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6384–6392.

[13] Stacey L Ivanko and Penny M Pexman. 2003. Context incongruity and irony processing. *Discourse processes* 35, 3 (2003), 241–279.

[14] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 757–762.

[15] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

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

[17] Jianing Li, Chaoqun Yang, Guanhua Ye, and Quoc Viet Hung Nguyen. 2024. Graph neural networks with deep mutual learning for designing multi-modal recommendation systems. *Information Sciences* 654 (2024), 119815.

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

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

[20] Chunmian Lin, Daxin Tian, Xuting Duan, Jianshan Zhou, Dezong Zhao, and Dongpu Cao. 2022. 3D-DFM: anchor-free multimodal 3-D object detection with dynamic fusion module for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems* 34, 12 (2022), 10812–10822.

[21] Xue Lin, Yushui Geng, Jing Zhao, Daquan Cheng, Xuefeng Zhang, and Hu Liang. 2023. Multi-modal medical image classification method combining graph convolution neural networks. In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 199–206.

[22] Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement.

In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 4995–5006.

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[24] Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li, and Jiafeng Guo. 2022. A Dual-Channel Framework for Sarcasm Recognition by Detecting Sentiment Conflict. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 1670–1680.

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

[26] Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu, and Hongfei Lin. 2023. Just like a human would, direct access to sarcasm augmented with potential result and reaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10172–10183.

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

[28] Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3930–3940.

[29] Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 9507–9515.

[30] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 704–714.

[31] Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*. 1136–1145.

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

[33] Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2468–2480.

[34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[35] Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng Chen. 2024. Gˆ 2SAM: Graph-Based Global Semantic Awareness Method for Multimodal Sarcasm Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 9151–9159.

[36] Changsong Wen, Guoli Jia, and Jufeng Yang. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2540–2550.

[37] Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023. DRIN: Dynamic Relation Interactive Network for Multimodal Entity Linking. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3599–3608.

[38] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference*. 2115–2124.

[39] Chujie Xu, Yong Du, Jingzi Wang, Wenjie Zheng, Tiejun Li, and Zhansheng Yuan. 2024. A joint hierarchical cross-attention graph convolutional network for multi-modal facial expression recognition. *Computational Intelligence* 40, 1 (2024), e12607.

[40] Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3777–3786.

[41] Zhe Yu, Di Jin, Xiaobao Wang, Yawen Li, Longbiao Wang, and Jianwu Dang. 2023. Commonsense knowledge enhanced sentiment dependency graph for sarcasm detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 2423–2431.

[42] Yufei Zeng, Zhixin Li, Zhenbin Chen, and Huifang Ma. 2024. A feature-based restoration dynamic interaction network for multimodal sentiment analysis. *Engineering Applications of Artificial Intelligence* 127 (2024), 107335.

[43] Duzhen Zhang, Feilong Chen, Jianlong Chang, Xiuyi Chen, and Qi Tian. 2023. Structure Aware Multi-Graph Network for Multi-Modal Emotion Recognition in Conversations. *IEEE Transactions on Multimedia* (2023).