Information Fusion 104 (2024) 102203

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Fact-sentiment incongruity combination network for multimodal sarcasm detection

Qiang Lu [a], Yunfei Long [b], Xia Sun [a,*], Jun Feng [a], Hao Zhang [c]

[a] School of Information Science and Technology, Northwest University, Xi'an 710127, China
[b] School of Computer Science and Electrical Engineering, University of Essex, Colchester CO43SQ, UK
[c] Graduate School, Shaanxi University of Chinese Medicine, Xianyang 712083, China

ARTICLE INFO

Keywords:
Multimodal sarcasm detection
Sarcasm incongruity
Dynamic connecting component
Cross-modal graph
Combination incongruity fusion

ABSTRACT

Multimodal sarcasm detection aims to identify whether the literal expression is contrary to the authentic attitude within multimodal data. Sarcasm incongruity method has been successfully applied to multimodal sarcasm detection, due to its ability to flexibly capture the intrinsic differences between modalities. However, previous incongruity methods primarily focused on the semantic level, often overlooking more specific forms of sarcasm incongruity. Sarcasm incongruity, in particular, encompasses fact incongruity, sentiment incongruity, and combination incongruity. Therefore, we propose a fact-sentiment incongruity combination network from a novel perspective, which draws the multimodal sarcastic relations by exploring the multimodal factual disparities, sentiment incongruity, and combination fusion. Specifically, we design a dynamic connecting component calculating dynamic routing probability weights via graph attention and mask routing matrices, which selects the most suitable image-text pairs to capture fact incongruity between images and text. Then, we retrieve sentiment relations between text tokens and image objects using external sentiment knowledge to reconstruct edge weights in the cross-modal graph matrix to capture sentiment incongruity. Furthermore, we introduce a combination incongruity fusion layer and cross-modal contrastive loss to fuse fact incongruity and sentiment incongruity for further enhancing the incongruity representations. Extensive experiments and further analyses on publicly available datasets demonstrate the superiority of our proposed model.

1. Introduction

Sarcasm is a distinct form of sentiment expression characterized by a contrast between the literal and implied meanings, typically conveying a scornful attitude that contradicts the user's true feelings [1–3]. As multimedia technologies evolve and online platforms grow, users increasingly express opinions using both text and image. This has amplified the importance of detecting sarcasm in multimodal datasets.

Multimodal sarcasm detection (MSD) is an emerging yet challenging task in natural language processing, and it aims to identify whether the literal expression is contrary to the authentic attitude by combining textual, visual, and other modalities [4–6]. Earlier studies used the principle of semantic incongruity for multimodal sarcasm detection. Some employed attention mechanisms and fusion strategies [7,8], while others leveraged pre-trained models for such modeling [9–11]. Additionally, some studies introduce the graph neural networks and external knowledge [12–14]. However, these incongruity methods mainly address the semantic level, frequently missing more nuanced forms of sarcasm incongruity [15,16]. Specifically, Liu et al. [15] consider that most existing studies only modeled the atomic-level inconsistencies between the text input and its accompanying image, ignoring more complex compositions for both modalities which have been proved to be effective in other related tasks, such as cross-modal retrieval [17] and image-sentence matching [18,19]. In addition, Wen et al. [16] focus on the inter-modal and dual incongruities. They discover that sarcasm incongruity not only involves semantic level but also includes attitude as a crucial factor [20].

In fact, sarcasm incongruity appears in more fine-grained forms, including fact, sentiment, and combination incongruities, as illustrated in Fig. 1. The textual description of Fig. 1(a) represents children go to school with wet roads and wind, while the actual situation in the image shows the children inside a school bus. The contrast between the words "wet roads" and "wind" with the image object "school bus" represents a complete contradiction between the text description and reality, illustrating factual incongruity. In Fig. 1(b), the word "yay" expresses joy for the snow, but the actual situation in the image is

* Corresponding author.
E-mail addresses: nwulq@stumail.nwu.edu.cn (Q. Lu), yl20051@essex.ac.uk (Y. Long), raindy@nwu.edu.cn (X. Sun), fengjun@nwu.edu.cn (J. Feng), 1271009@sntcm.edu.cn (H. Zhang).

https://doi.org/10.1016/j.inffus.2023.102203
Received 25 October 2023; Received in revised form 22 November 2023; Accepted 15 December 2023
Available online 18 December 2023

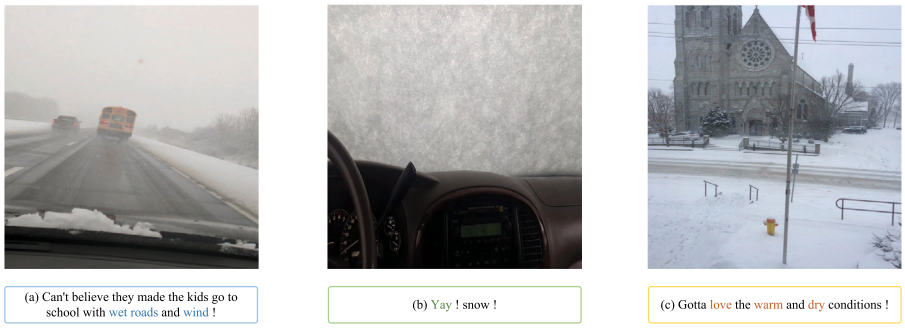

Fig. 1. Examples of Twitter data with multimodal sarcasm.

that the snow covers the entire windshield, making it impossible to drive the car. The wording contrasts with the actual sentiment attitude, representing sentiment incongruity. In Fig. 1(c), the text description of "warm" and "dry" strongly contrasts with the image object "snow", and the word "love" further intensifies the level of sarcasm through a false sentiment attitude, which combines factual incongruity and sentiment incongruity.

In this paper, we introduce the Fact-Sentiment Incongruity Combination Network (FSICN) from a new angle, capturing multimodal sarcasm by examining fact, sentiment, and combination incongruities. Specifically, we design a fact incongruity module containing dynamic connecting component, which calculates dynamic routing probability weights via graph attention mechanism and mask routing matrices to select the most suitable image-text pairs to capture fact incongruity. Then, we construct a sentiment incongruity module by retrieving sentiment relations between text tokens and image objects, and introduce the external sentiment knowledge into the cross-modal graph matrix to capture sentiment incongruity. Furthermore, we utilize a combination incongruity fusion to fuse fact incongruity and sentiment incongruity to capture composite incongruity, and introduce a cross-modal contrastive loss to further enhance the multimodal incongruity representations. Experiments on the publicly-available dataset show our proposed FSICN outperforms the baselines that rely solely on text or image by achieving a 22.7% and 6.7% improvement in accuracy, respectively. In addition, FSICN outperforms advanced multimodal baselines, achieving improvements of 0.96% in accuracy, 2.55% in binary-average F1-score, and 1.17% in macro-average F1-score.

The main contributions of our paper can be summarized as follows:

- We propose a fact-sentiment incongruity combination network from a novel perspective to capture the fine-gained sarcasm incongruity.
- We design a dynamic connecting component that calculates dynamic routing probability weights to adaptively select the most suitable image-text pairs to capture fact incongruity.
- We reconstruct a cross-modal graph by retrieving sentiment relations between text tokens and image objects using external sentiment knowledge to capture sentiment incongruity.
- Experimental results on the publicly-available datasets MSD illustrate that our proposed model outperforms advanced baseline methods and demonstrate the superiority of our model.

The rest of this paper is organized as follows. After introducing related works in Section 2, we propose a fact-sentiment incongruity fusion network in Section 3. Then we report the experimental details and conduct a detailed experimental analysis in Section 4. Finally, we summarize our work and provide a direction of future work in Section 5.

2. Related work

Previous studies mostly focused on using textual information for sarcasm detection [21–23]. With the development of multimedia technology and the popularity of multimodal information, multimodal sarcasm detection has attracted considerable attention in recent years [11,13,16]. Unlike textual sarcasm detection, multimodal sarcasm detection seeks to determine if the literal expression contradicts the genuine attitudes across various modalities. In this section, we will discuss related work in two parts: multimodal sarcasm detection and multimodal sentiment analysis.

2.1. Multimodal sarcasm detection

Early studies merged text and image data to detect multimodal sarcasm. Schifanella et al. [24] applied two methods for multimodal sarcasm detection: one combined visual semantics with text features from an external dataset, while the other used pre-trained visual neural networks from ImageNet [25]. Caiet et al. [7] adopted a hierarchical fusion approach from a hierarchical fusion perspective. They used bidirectional long-short memory network (LSTM) [26] to extract text features, combining them with image and image attribute features to reconstruct representation vectors and weighted averages to detect sarcasm.

However, previous fusion methods fail to capture the multimodal interaction and implicit sarcasm relations. Consequently, recent studies attempt to model multimodal sarcasm by exploring the implicit incongruities among different modalities. Xu et al. [8] modeled the semantic associations in cross-modal contexts by decomposing the network to represent the commonalities and differences between images and text. Drawing inspiration from the self-attention mechanisms, Pan et al. [27] designed a cross-modal attention mechanism based on BERT [28] to capture intra-modal and inter-modal incongruities in multimodal sarcasm detection, and they also applied a shared attention mechanism to model contradictions within the text. Due to the incongruity is a crucial clue in determining sarcasm, Liang et al. [12] designed an interactive Graph Convolutional Network to establish sarcasm incongruity by creating heterogeneous intra-modal and cross-modal graphs for each multimodal input. Previous approaches overlooked the wealth of information contained in external knowledge. Therefore, Liu et al. [15] proposed a hierarchical framework based on a multi-head cross-attention mechanism and graph neural networks, which integrates various knowledge resources for detecting sarcasm incongruity. Yue et al. [14] identified prior knowledge and cross-modal semantic contrast as essential factors in sarcasm detection, and introduced a novel model that incorporated prior knowledge from the ConceptNet knowledge base and incorporated contrastive learning to enhance the spatial distribution of samples.

A few incongruity methods has been successfully applied to multimodal sarcasm detection, due to its ability to flexibly capture the intrinsic differences between modalities [29–31]. However, these methods primarily focused on the semantic level, often overlooking more specific forms of sarcasm incongruity. Sarcasm incongruity, in particular, encompasses fact incongruity, sentiment incongruity, and combination incongruity. Therefore, from a unique perspective, we introduce a fact-sentiment incongruity combination network (FSICN). This

network depicts multimodal sarcastic relations by probing into the multimodal factual discrepancies, sentiment incongruities, and their integrated fusion.

2.2. Multimodal sentiment analysis

Recent years have seen growing interest in multimodal sentiment analysis [32–35], which is closely tied to sarcasm detection. Sarcasm, a unique sentiment expression, subtly conveys dissatisfaction in a positive way. Given this relationship, detecting sentiment in multimodal data is vital for accurate sarcasm detection.

Earlier research typically used deep neural networks like convolutional neural networks (CNNs) [36], long-short term memory networks (LSTMs) [37], Memory networks [38], pre-trained models (PTMs) [39], and graph convolutional networks (GCNs) [40] for multimodal sentiment prediction. CNN-based methods demonstrate proficiency in capturing local features. Poria et al. [41] introduced a multi-kernel learning method that utilized CNN to extract textual and visual features for multimodal sentiment analysis. Because CNNs have limitations in capturing global multimodal features, Chen et al. [42] constructed a time attention mechanism on top of LSTM to achieve finer modal fusion for multimodal sentiment analysis. However, this approach neglected the semantic sentiment information conveyed by words, Zhu et al. [43] proposed an sentiment knowledge enhanced attention fusion network that incorporated additional sentiment knowledge representations from external knowledge bases.

Due to the ability to extract extra knowledge from large-scale datasets, pre-trained models have demonstrated outstanding performance in constructing multimodal representations. Ye et al. [44] designed a cross-modal contrastive learning method based on pre-trained models and introduced a sentiment-aware pre-training objective for multimodal sentiment analysis. Existing studies relies on cascade operations for feature fusion, overlooking the deep interactions between different modalities. To address this issue, Liu et al. [45] proposed a modality translation module to construct missing joint features. Under the supervision of pre-trained models, it generated joint features for the uncertain missing modality to facilitate multimodal sentiment prediction. GCNs enable the learning of feature representations in graph data, automatically capturing relations between data and effectively integrating heterogeneous data. Huang et al. [46] proposed a temporal graph convolutional network that leveraged modality-specific graph learning to embed nodes with underlying sequential semantics of discourse for multimodal sentiment prediction. Due to the neglect of fine-grained multimodal information in existing methods, Wang et al. [47] employed text-image pairs and graph structures to explore both global and local fine-grained sentiment details for multimodal sentiment analysis.

Though previous methods use diverse techniques for multimodal sentiment analysis, they miss capturing cross-modal graph dependencies and thus cannot precisely identify sentiment cues. Hence, we introduced a cross-modal graph convolutional network to address this and enhance multimodal sarcasm detection.

3. Methodology

In this section, the proposed **F**act-**s**entiment **i**ncongruity **c**ombination **n**etwork (FSICN) is described in detail. As demonstrated in Fig. 2, the architecture of FSICN contains five components: (1) Text and image encoding module. (2) cross-modal interactive module (CIM). (3) Fact incongruity module (FIM). (4) Sentiment incongruity module (SIM). (5) Combination fusion module (CFM). First, the text and image encoding module is applied to capture the general features via BERT [28] and ViT [48] pre-trained models. Then, cross-modal interactive module is designed to build cross-modal interactive relations between image and text. Next, fact incongruity module calculates dynamic routing probability weights using the masked matrix to obtain the most suitable image-text pairs adaptively, and sentiment incongruity module constructs the cross-modal graph by integrating sentiment clues and calculating the semantic similarity between text words and image patches. Finally, the obtained fact incongruity and sentiment incongruity embeddings are fused into combination fusion layer via activation function to predict multimodal sarcasm.

3.1. Task and notation definition

The task of MSD can be formulated as follows: given a image-text pair which the text contains m words and the image is split into n patches. The image-text pair is denoted as $Text = \{e_i^t \mid 1 \le i \le m\}$ and $Image = \{e_j^v \mid 1 \le j \le n\}$, where e_i^t denotes the ith word of given sentence, and e_j^v denotes the jth patch of the corresponding image. The goal of MSD is to learn a classifier to predict the sarcasm label $y \in \{sarcasm, non-sarcasm\}$ for each image-text pair.

3.2. Text and image encoding

To obtain the general text features for model training, we use the pre-trained BERT [28] which has acquired knowledge from the large-scale datasets as the text encoder. Similarly, the image features are extracted via pre-trained Vision Transformer (ViT) in lines with other baselines [48], which has achieved excellent performance in image encoding. The text and image embeddings extraction can be formulated as:

$$\begin{aligned} T &= [t_1, t_2, \dots, t_m] = BERT(Text) \\ V &= [v_1, v_2, \dots, v_n] = ViT(Image) \end{aligned} \tag{1}$$

where $T \in \mathbb{R}^{m \times d_t}, V \in \mathbb{R}^{n \times d_v}$ represents the text and image embeddings. $t_i \in \mathbb{R}^{d_t}$ is the text embedding of the $i-th$ word, and $v_i \in \mathbb{R}^{d_v}$ is the image embedding of the jth patch in the image. d_t, d_v denote the dimension of text and image embedding.

3.3. Cross-modal interactive module

The interactive relation between text and image in multimodal sarcasm is crucial for recognizing sarcastic content. Text provides the linguistic aspect of irony or exaggeration, while image offers visual elements such as expressions, actions, or scenes that can further emphasize or explain the meaning of the text, aiding in understanding the sentiment and contextual aspects of sarcasm. Therefore, we first map the text and image features into the joint multimodal space, and calculate the similarity relation with L2-normalization to build the interaction between sarcastic images and text.

$$\begin{aligned} L^t &= LN(T \cdot W_t) \\ L^v &= LN(V \cdot W_v) \\ E &= (L^t \cdot (L^v)^T) * e^t \end{aligned} \tag{2}$$

where $L^t \in \mathbb{R}^{m \times d_e}, L^v \in \mathbb{R}^{n \times d_e}$ represents the joint multimodal embeddings. $W_t \in \mathbb{R}^{d_t \times d_e}, W_v \in \mathbb{R}^{d_v \times d_e}$ are learnable weight. $E \in \mathbb{R}^{m \times n}$ denotes the interactive matrix, and e^t is the learned temperature parameter. After that, we integrate the interactive matrix with text and image features to fully leverage the complementary information from different modalities.

$$\begin{aligned} T^c &= [t_1^c, t_2^c, \dots, t_m^c] = \frac{\exp(E_t)}{\sum_{i=1}^{m} \exp(E_i)} * T \\ V^c &= [v_1^c, v_2^c, \dots, v_n^c] = \frac{\exp(E_t)}{\sum_{j=1}^{n} \exp(E_j)} * V \end{aligned} \tag{3}$$

where $T^c \in \mathbb{R}^{m \times d_t}, V^c \in \mathbb{R}^{n \times d_v}$ represent the cross interactive representations which text towards image and image towards text. $exp()$ is the exponential function, and $E_t \in \mathbb{R}^{m \times n}$ represents the interactive value of the tth token in text and image. Thus far, we have achieved semantic interaction and information complementarity between text and images via the cross-modal interaction module, providing the enhanced cross-modal embeddings for the subsequent fact incongruity module and sentiment incongruity module.

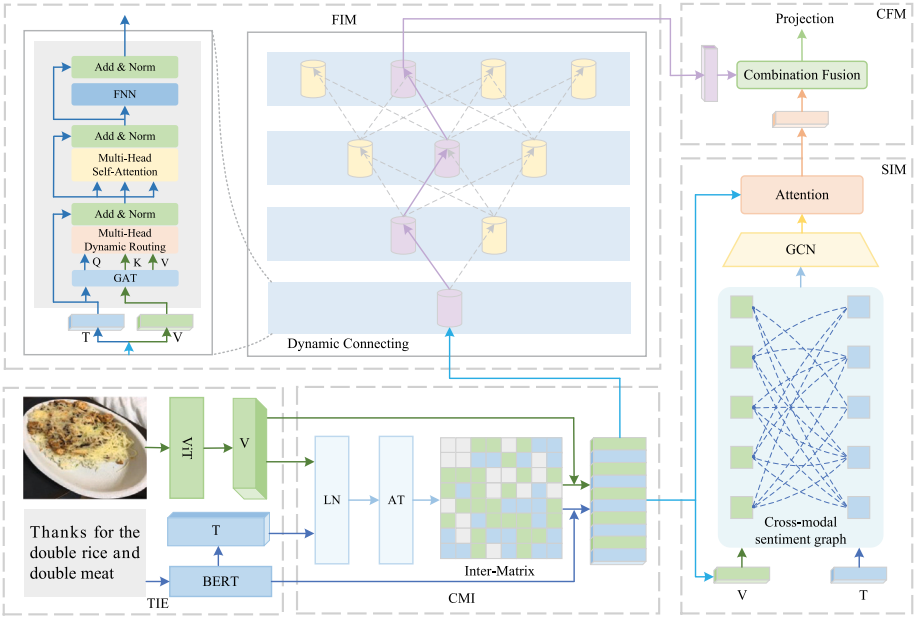

Fig. 2. The architecture of proposed FSICN contains five components: TIE for great features extraction. CMI is cross-modal interactive module used to capture interactive relations between text and image. FIM and SIM are fact incongruity and sentiment incongruity modules to extract the fact and sentiment incongruity embeddings. CFM is combination incongruity fusion to fuse fact and sentiment incongruity embeddings to predict sarcasm.

3.4. Fact incongruity module

Fact describes the existence of objects or entities, and is perceived through semantic information. Multimodal fact incongruity refers to the discrepancies between information or facts in multimodal data. These discrepancies are manifested as objects described in text that either do not exist in real images or do not match them. The meanings of words or phrases in the text may contrast with the objects depicted in the visual data. By modeling fact incongruity in multimodal sarcasm data, this enables us better distinguish false and fact information from both global and local perspectives.

Inspired by previous studies [49,50], we have developed a dynamic connecting (DC) component. This component is designed to dynamically capture fact incongruity between images and text by selecting the most suitable module based on different image-text pairs. The DC component comprises several key elements, including graph attention network (GAT) [51], multi-head dynamic routing layer (DynRT) [49], multi-head self-attention (MHA) and feed-forward network (FNN). The DC component can be calculated as follows:

$$\begin{aligned} F_k &= LN(FNN(F^a_{k-1}) + F^a_{k-1}) \\ F^a_{k-1} &= LN(MHA(F^d_{k-1}) + F^d_{k-1}) \end{aligned} \tag{4}$$

Where $F_k \in \mathbb{R}^{m\times d_f}$ is the output of kth DC layer which represent the fact incongruity embeddings. $F^a_{k-1} \in \mathbb{R}^{m\times d_t}$ refers to the MHA embeddings, and $F^d_{k-1} \in \mathbb{R}^{m\times d_t}$ denotes the $(k-1)$-th DynRT embeddings, among $d_t = d_v = d_f$.

The MHA mechanism receives input from the DynRT layer, and utilize the DynRT to calculates dynamic routing probability weights using the masked matrix, enabling the adaptive selection of the most suitable image-text pairs. The DynRT layer is as follows:

$$F^d_{k-1} = Concat(head_1, head_2, \dots, head_h)W_O \tag{5}$$

Where $F^d_{k-1} \in \mathbb{R}^{m\times d_t}$ denotes the $(k-1)$-th DynRT embeddings. $Concat$ is the concatenation operation. $W_O \in \mathbb{R}^{d_t\times d_t}$ is the parameter matrix. $head_i \in \mathbb{R}^{m\times d_r}$ is calculated by routing function for each image-text pair, and hidden dimension $d_t = h * d_r$. The $head_i$ can be calculated as follows:

$$head_i = softmax(\frac{Q_T K_V{}^{\mathrm{T}}}{\sqrt{d_k}} \otimes \sum_{i=1}^{k} \alpha_i A_i)V_V \tag{6}$$

Where $Q_T = T^g W_Q \in \mathbb{R}^{m\times d_r}$, $K_V = V^g W_K \in \mathbb{R}^{n\times d_r}$, $V_V = V^g W_V \in \mathbb{R}^{n\times d_r}$ represent query, key and value via linear transformation, and $W_Q \in \mathbb{R}^{d_t\times d_r}, W_K \in \mathbb{R}^{d_v\times d_r}, W_V \in \mathbb{R}^{d_v\times d_r}$ are parameter matrices. $A_i \in \mathbb{R}^{m\times n}$ denotes masked matrixh to calculate the coupling coefficents between each image-pair pair. If the image patch within the attention span of the text target, the value of dynamic mask matrix is set to 1, otherwise set to 0. $\alpha_i = softmax(MLP(AttentionPool(V_g))) \in \mathbb{R}^k$ refers to the routing probability weight. $AttentionPool()$ is the adaptive average pooling method, and $MLP()$ is the multi-layer perception. $T^g \in \mathbb{R}^{m\times d_t}, V^g \in \mathbb{R}^{m\times d_t}$ represent the graph attention embeddings towards text and image to establish text semantic and graph edge relations, and can be calculated as follows:

$$\begin{aligned} T^g &= \frac{\exp(\sigma(a^t[t^c_i \cdot W_T \| t^c_j \cdot W_T]))}{\sum_{i=1}^{m}\exp(\sigma(a^t[t^c_i \cdot W_T \| t^c_k \cdot W_T]))} * T^c \\ V^g &= \frac{\exp(\sigma(a^v[v^c_i \cdot W_V \| v^c_j \cdot W_V]))}{\sum_{i=1}^{n}\exp(\sigma(a^v[v^c_i \cdot W_V \| v^c_k \cdot W_V]))} * V^c \end{aligned} \tag{7}$$

Where σ denotes the $LeakyReLU$ activation function. $a^t \in \mathbb{R}^{2d_t}, a^v \in \mathbb{R}^{2d_v}$ and $W_T \in \mathbb{R}^{d_t\times d_t}, W_V \in \mathbb{R}^{d_v\times d_v}$ are the learnable parameter.

3.5. Sentiment incongruity module

Sarcasm often involves implicitly expressing dissatisfaction in a positive manner while conveying sentiments or attitudes. For example, in the sentence "what a wonderful weather!" which corresponds to with an image of a rainy day. The word "wonderful" in the text conveys a highly positive sentiment, creating a strong sentiment inconsistency with the negative sentiment implied by the rainy weather in the image. Hence, sarcasm implies sentiment, making the sentiment detection in multimodal data a crucial factor in accurately predicting sarcasm.

To extract sentiment information from multimodal sarcasm, we approach the extraction of sentiment incongruity as a multimodal sentiment analysis task. Giving the intricate associations and dependencies in multimodal sentiment, we utilize graph convolutional networks (GCNs) to integrate features from various modalities by propagating information across the graph, which allows model to capture cross-modal sentiment relations. We first construct the textual graph based

on dependency tree.[1]

$$D^T_{i,j} = \begin{cases} 1, & \text{if } i = j \text{ or } \{t^c_i, t^c_j\} \text{ in} \in \mathrm{T} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Where $D^T_{i,j} \in \mathbb{R}^{m\times m}$ is the adjacency matrix of textual modality, and t^c_i, t^c_j denote the words of sentence. In the adjacency matrix, each node is set to be adjacent to itself, and the value of diagonal is all set to one. Next, we build the visual graph as follow.

$$D^V_{i,j} = \begin{cases} 1, & \text{if } i = j \text{ or } \{t^c_i, t^c_j\} \in \mathrm{R} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Where $D^V_{i,j} \in \mathbb{R}^{n\times n}$ is the adjacency matrix of visual modality, and v^c_i, v^c_j denote the image patches. After obtaining the textual and visual graph, we feed graphs and cross interactive representations into GCN layers to generate the graph representation.

$$\begin{aligned} g^t_i &= ReLU(\sum_{j=1}^{m} D^T\ W_L\ g^{t-1}_j + b_L) \\ g^v_i &= ReLU(\sum_{j=1}^{n} D^V\ W_P\ g^{v-1}_j + b_P) \end{aligned} \tag{10}$$

$$\begin{aligned} H^t &= \sum_{k=1}^{m} \eta^t_k t^c_i, \eta^t_k = \frac{\exp(\beta^t_k)}{\sum_{i=1}^{m}\exp(\beta^t_i)}, \beta^t_k = \sum_{i=1}^{m} g^t_i t^c_k \\ H^v &= \sum_{k=1}^{n} \eta^v_k v^c_i, \eta^v_k = \frac{\exp(\beta^v_k)}{\sum_{i=1}^{n}\exp(\beta^v_i)}, \beta^v_k = \sum_{i=1}^{n} g^v_i v^c_k \end{aligned} \tag{11}$$

Where $g^t_i \in \mathbb{R}^{d_t}, g^v_i \in \mathbb{R}^{d_v}$ represent the graph hidden representations, and $g^{t-1}_j \in \mathbb{R}^{d_t}, g^{v-1}_j \in \mathbb{R}^{d_v}$ denote the representation evolved from the preceding GCN layer. $W_L \in \mathbb{R}^{d_t\times d_t}, W_P \in \mathbb{R}^{d_v\times d_v}$ and $b_L \in \mathbb{R}^{d_t}, b_P \in \mathbb{R}^{d_v}$ are the weight parameters. $ReLU()$ is a nonlinear activation function. $H^t \in \mathbb{R}^{m\times d_t}, H^v \in \mathbb{R}^{n\times d_v}$ is the final graph embeddings.

Then, since the weights of the edges are important in graph information aggregation [13,52,53], we construct a cross-modal graph by integrating sentiment clues and calculating the semantic similarity between text words and image patches.

$$D^{cross}_{i,j} = \begin{cases} 1 + sim(t^c_i, v^c_i) e^{-\delta(t^c_i)\delta(a^c_j)}, & \text{if } i < m, j \geq n \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Where $D^{cross}_{i,j} \in \mathbb{R}^{(m+n)\times(m+n)}$ is the corss-modal graph representation. sim refer to the similarity calculation. t^c_i, a^c_j denote the word of sentence and attribute of image patch. $\delta(t^c_i) \in [-1, 1]$ is the sentiment weight of word t^c_i in SenticNet [54]. Words found in SenticNet are assigned their corresponding values, while others are set to 0.

Finally, we create the cross-modal sentiment embeddings at the top of graph representations H^t and H^v via graph convolution operation.

$$g^c_i = ReLU(\sum_{j=1}^{m+n} D^{cross}\ W_C\ g^{c-1}_j + b_C) \tag{13}$$

$$H^c = \sum_{t=1}^{m+n} \eta^c_t h^c_t, \eta^c_t = \frac{\exp(\beta^c_t)}{\sum_{i=1}^{m+n}\exp(\beta^c_i)}, \beta^c_t = \sum_{i=1}^{m+n} g^c_i h^c_t \tag{14}$$

Where $H^c \in \mathbb{R}^{d_f}$ is the finally sentiment incongruity embeddings. $g^c_0 = [H^t, H^v] = \{h^c_1, h^c_2, \ldots, h^c_{n+m}\} = \{h^t_1, h^t_2, \ldots, h^t_m, h^v_1, h^v_2, \ldots, h^v_n\}$.

3.6. Combination incongruity fusion

Utilizing the fact incongruity module and sentiment incongruity module, we input embeddings F^a and H^c into combination incongruity module to generate combination embeddings to predict sarcasm.

$$\begin{aligned} F &= Mean(F^a) \\ S &= Mean(H^c) \end{aligned} \tag{15}$$

$$y^c = softmax(W_y(LN(F \cdot S)) + b_y) \tag{16}$$

Where $Mean(.)$ is the average function. $F, S \in \mathbb{R}^{d_f}$ represent global embeddings of fact incongruity and sentiment incongruity. y^c is the predicted probability of all the possible labels, and $W_y \in \mathbb{R}^{d\times d}$ and $b_y \in \mathbb{R}^d$ are trainable parameters.

3.7. Optimization objectives

For our FSICN model, the overall learning of the model is to optimize all the parameters, and minimize the loss function as far as possible. The overall loss is as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{mse} + \mathcal{L}_{cl} \tag{17}$$

Where $\mathcal{L}_{ce}$ is the cross-entropy loss, $\mathcal{L}_{mse}$ is the Mean Square Error loss. To further enhance the image sentiment representations, we introduce the graph contrastive learning strategy $\mathcal{L}_{cl}$. Specifically, we use graph convolution operation to extract the text and image graph representations H^t and H^v. For image-text pairs, a substantial difference in sentiment polarity implies that text and image embeddings should be correspondingly pushed apart. Otherwise, they should be pulled closer together. Therefore, we utilize the Kullback-Leible (KL) divergence to calculate the graph contrastive learning.

$$\begin{aligned} \mathcal{L}_{ce} &= -\frac{1}{N}\sum_{i}^{N} y^c_i \log\left(\hat{y}^c_i\right) \\ \mathcal{L}_{mse} &= \frac{1}{N}\sum_{i=1}^{N} \|y^s_i - \hat{y}^s_i\|^2 \\ \mathcal{L}_{cl} &= \frac{1}{2} D_{KL}(H^t \| H^v) + \frac{1}{2} D_{KL}(H^v \| H^t) \end{aligned} \tag{18}$$

Where y^s_i is the predicted probability of sentiment incongruity embeddings h^c via softmax function.

4. Experiments

In this section, we initially describe the experimental datasets in Section 4.1, followed by the implementation details and baseline models in Sections 4.2 and 4.3. To evaluate the performance of proposed model, we compare it with advanced baselines on MSD, and utilize the Accuracy (shorten as Acc), Binary F1-score, and Macro F1-score as the evaluation metrics in Section 4.4. Next, in Section 4.5, we conduct an ablation study to analyze the contribution of CMI, FIM, SIM and CFM module. We also explore the influence of the number of DC layers and cross-modal GCN layers to model performance in Sections Section 4.6. Finally, we visualize the dynamic connection layer in FIM and the cross-modal graph matrix in EIM to gain a deeper understanding of the principles underlying FIM and SIM in Section 4.7.

4.1. Experimental datasets

We conduct experiments on publicly available benchmark sarcasm datasets in line with the most of state-of-the-art works in this area: Multimodal Sarcasm Detection (MSD) Dataset [7]. This dataset collects image-text pairs containing some specific hashtag (e.g., #sarcasm, etc.) as sarcastic examples from Twitter,[2] and collects image-text pairs without such hashtags as non-sarcastic examples, as shown in Tabel 1. To improve the quality of the dataset, Cai et al. [7] discards tweets containing $sarcasm, sarcastic, irony, ironic$ as regular words and discards $URLs$, and randomly divides dataset into the training set, validation set, and testing set with the ratio of 80%, 10%, and 10%. Consistent with previous studies, we evaluate our model using standard metrics, including accuracy, precision, recall, binary-average, and macro-average results.

[1] We use spaCy toolkit to construct the dependency tree: https://spacy.io/

[2] https://twitter.com/home

Table 1
Statistics of the MSD data.

	Training	Validation	Testing
Sarcastic	8642	959	959
Non-Sarcastic	11 174	1451	1450
All	19 816	2410	2409

For a more comprehensive validation, we conduct experiments on a multimodal datasets that contain the sentiment, emotion and sarcasm labels, called Memotion dataset originates from SemEvel 2020 Task 8 [55]. Memotion contains 6992 samples which consists of 5449 sarcastic samples and 1543 non-sarcastic samples. Each memo data point has been labeled with semantic dimensions, e.g., sentiment and type of emotion, e.g., sarcasm, humor, etc. The speaker identifiers of all the utterances are also recorded. Consistent with previous studies, we evaluate our model using standard metrics, including accuracy, precision, recall, binary-average.

In addition, we also conduct experiments on another multimodal meme dataset, called MultiBully annotated with bully, sentiment, emotion and sarcasm labels [56]. The MultiBully dataset comprises 5854 samples, divided into 2233 sarcastic samples and 3641 non-sarcastic samples, collected from open-source Twitter and Reddit platforms. The dataset also consists of two modalities, text and image. Consistent with previous studies, we evaluate our model using standard metrics, including accuracy, precision, recall, binary-average.

4.2. Implementation details

For a fair comparison, we follow [7] to process the MSD dataset. In our experiments, we adopt the pre-trained uncased BERT [28] with 768 dimension to initialize the text embedding, and the image embedding is obtained by the pre-trained ViT [48] with 768 dimension. The image is split into 49 (7*7) patches and the resolution of visual region patch is set to 32. We set the number of DC layers to 4 and the number of GCN layers to 2. The hidden state dimension is configured as 512, and the output hidden state is set to 768. For optimization, we employ the Adam optimizer with a learning rate of 0.00002 across all models. To optimize the model training, we average the experimental results of 20 runs with random initialization, and use early-stopping with patience value of 5.

4.3. Baseline models

To assess the performance of FSICN, we compare with a series of state-of-the-art baselines, summarized as image-modality methods, text-modality methods and multimodal methods.

Image-modality methods. These baselines use visual information for sarcasm detection are as follows:

- **Image** [7] utilizes the visual representation via ResNet for sarcasm detection.
- **ViT** [48] uses the [CLS] token representation as the input of pre-trained visual model to detect the sarcasm.

Text-modality methods. These baselines use textual information for sarcasm detection are as follows:

- **Bi-LSTM** [57] uses a bidirectional long-short memory network for text classification.
- **TextCNN** [58] utilizes a convolutional neural network for text classification.
- **SIARN** [59] proposes an attention-based neural model to explicitly model contrast and incongruity for sarcasm.
- **SMSD** [60] designs a self-matching network to capture sarcasm incongruity information by exploring interactions between different words.
- **BERT** [28] is a pre-trained uncased model which takes [CLS] text [SEP] as input for text classification.
- **ALBERT** [61] presents two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT.
- **XLNet** [62] enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order, and overcomes the limitations of BERT thanks to its autoregressive formulation.
- **RoBERTa** [63] presents a replication study of BERT pretraining that carefully measures the impact of many key hyperparameters and training data size.

Multimodal methods. These baselines take both textual and visual information as input to detect sarcasm are as follows:

- **HFM** [7] proposes a multimodal hierarchical fusion model to detect sarcasm.
- **D&R Net** [8] designs a decomposition and relation network by modeling cross-modality contrast and semantic association for sarcasm detection.
- **Res-BERT** [27] uses the BERT to encode text and combine text and image features for sarcasm prediction.
- **Att-BERT** [27] constructs the attention mechanism to construct the inter-modality attention to capture inter-modality incongruity.
- **UPB-MTL** [64] uses ALBERT to represent the textual utterance and uses VGG-16 to represent the accompanying image.
- **FAT-MTL** [65] designs a genetic program tree to predict the inter-task covariance matrix.
- **A-MTL** [66] proposes Ie-Attention and Ia-Attention to learn the relation between different segments and the relation within the same segment.
- **RCNN-RoBERTa** [67] utilizes pretrained RoBERTa vectors to represent the utterance and uses an RCNN to obtain its contextual representation.
- **InCrossMGs** [12] designs the heterogeneous in-modal and cross-modal graphs via graph convolutional network to detect sarcasm.
- **CMGCN** [13] constructs a cross-modal graph convolutional network to draw the sarcasm reltions for sarcasm prediction.
- **HKEmodel** [15] combines the atomic-level congruity and atomic-level congruity based on graph convolutional network to detect sarcasm.
- **MILNet** [11] designs the local semantic-guided and global incongruity learning modules for sarcasm detection.
- **DIP** [16] proposes a dual Incongruity perceiving network which used the leverage gaussian distribution and contrastive learning for sarcasm detection.
- **DynRT** [49] uses the hierarchical co-attention to construct the dynamic path for detecting thg cross-modal sarcasm incongruity.
- **KnowleNet** [14] incorporates prior knowledge via the ConceptNet knowledge and captured the cross-modal semantic similarity for sarcasm prediction.
- **VisualBERT** [68] is a simple and flexible framework for modeling a broad range of vision-and-language tasks.
- **ViLBERT** [69] is a model for learning task-agnostic joint representations of image content and natural language.

4.4. Results and analysis

To evaluate the performance of proposed model, we utilize Acc, Binary-Average F1-score, and Macro-Average F1-score on MSD dataset, as shown in Table 2. The results demonstrate that FSICN outperforms all the state-of-the-art baselines.

Firstly, our proposed FSICN, which combines both text and image modalities, outperforms the baselines that rely solely on text or image

Table 2
Performance of FSICN compared to state-of-the-art baselines on MSD with the evaluation metrics acc, binary-average F1-score and macro-average F1-score.

Modality	Model	Acc	Binary-Average			Macro-Average		
			P↑	R↑	F1↑	P↑	R ↑	F1↑
Image	Image	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.68	71.35	68.40
Text	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	TextCNN	80.03	74.29	76.39	75.32	78.03	78.28	78.15
	SIARN	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
Image + Text	HFM	86.63	83.84	84.18	84.01	86.24	86.28	86.26
	D&R Net	84.02	77.97	83.42	80.60	–	–	–
	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
	CMGCN	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	HKEmodel	87.36	81.84	86.48	84.09	–	–	–
	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
	KnowleNet	88.87	88.59	84.18	86.33	88.83	88.21	88.51
Ours	FSICN	**90.55**	**89.93**	**89.51**	**89.72**	**90.16**	**90.42**	**90.29**

by achieving a 22.7% and 6.7% improvement in accuracy, respectively. This highlights the importance of leveraging both modalities for more accurate sarcasm detection. Additionally, we observed that the text-only baseline significantly outperforms the image-only baseline, indicating that textual semantics play a more crucial role in sarcasm detection compared to images.

Secondly, FSICN outperforms advanced multimodal baselines, achieving improvements of 0.96% in accuracy, 2.55% in binary-average F1-score, and 1.17% in macro-average F1-score. We attribute the improvement to the analysis of baseline structures, and the primary reasons are as follows: On one hand, the HFM and D&R Net models are among the earliest researches in multimodal sarcasm detection, which utilize the ResNet+LSTM to extract features from both image and text modalities. Since sarcasm clues in the text are more crucial, the performance of HFM and D&R Net is not satisfactory when compared to the baseline using the BERT model. On the other hand, models such as Res-BERT, IncrossMGs, CMGCN, HKE, MILNet, DIP, and KnowleNet capture the implicit incongruity between text and images from various perspectives to detect sarcasm. However, these models lack fine-grained analysis of incongruity and the ability to dynamically capture semantic correlations between text and images as well as cross-modal sentiment interactions. The proposed FSICN takes a more fine-grained approach by analyzing incongruity in multimodal sarcasm from the perspectives of factual incongruity, sentiment incongruity, and combination incongruity. FSICN utilizes dynamic connections, cross-modal sentiment graph convolution, and combination fusion to establish dynamic semantic correlations, cross-modal sentiment interactions, and consistent representations, enhancing the accuracy of multimodal sarcasm detection.

Recently, Transformer-based pre-trained models have demonstrated powerful performance. Therefore, building upon the BERT pre-trained model, we test multimodal models based on both unimodal and multimodal Transformers on the MSD dataset, as shown in Table 3. Firstly, compared to the BERT baseline model, RoBERTa has achieved an increase of 4.43% in accuracy and 5.67% in F1-score. Secondly, in the multimodal scenario, we replaced the BERT baseline in the KnowleNet and DynRT models with ALBERT and RoBERTa. We observe that the performance of the substituted models far surpassed the BERT baseline. Finally, building upon our proposed FSICN model, we replace the BERT baseline with XLNet, ALBERT, and RoBERTa, respectively. We observe that the performances of XLNet, ALBERT, and RoBERTa outperform the BERT baseline, with RoBERTa exhibiting the best performance. Specifically, using RoBERTa resulted in a performance improvement of

Table 3
bluePerformance of FSICN compared to other transformer-based baselines on MSD with the evaluation metrics Acc, Binary-average F1-score.

Modality	Model	Acc	Binary-Average		
			P↑	R↑	F1↑
Text	BERT	83.85	78.72	82.27	80.22
	RoBERTa	88.28	86.32	85.48	85.89
Image + Text	VisualBERT	83.51	76.66	82.94	79.68
	ViLBERT	84.68	77.52	86.37	81.71
	KnowleNet + BERT	88.87	88.59	84.18	86.33
	KnowleNet + ALBERT	92.69	91.57	90.85	91.21
	DynRT-Net + BERT	89.77	–	–	87.36
	DynRT-Net + RoBERTa	93.49	–	–	93.21
Ours	FSICN + BERT (110M)	90.55	89.93	89.51	89.72
	FSICN + XLNet (110M)	92.53	92.12	90.96	91.54
	FSICN + ALBERT (125M)	93.17	92.83	91.64	92.23
	FSICN + RoBERTa (125M)	**94.71**	**93.62**	**93.28**	**93.45**

Table 4
bluePerformance of FSICN compared to the baselines on Memotion dataset with the Precision, Recall and Binary-average F1-score.

Modality	Model	P	R	F1
Image + Text	RCNN-RoBERTa	50.44	50.77	50.52
	UPB-MTL	51.38	51.71	51.59
	FAT-MTL	43.54	44.21	43.89
	A-MTL	60.23	59.74	59.85
	HFM	44.43	44.68	44.59
	CMGCN	61.39	61.95	61.67
	HKEmodel	61.32	61.47	61.41
	MILNET	62.21	61.77	61.99
	DIP	62.43	61.76	62.09
Ours	FSICN	**63.57**	**62.88**	**63.22**

4.16% in accuracy and 3.73% in F1-score compared to using BERT. We consider this phenomenon to the following reasons: XLNet uses the same 110M parameter size as BERT and incorporates architecture of BERT into the Transformer-XL model, overcoming the shortcomings of the BERT model and thus achieving performance improvement. ALBERT and RoBERTa use larger 125M parameter sizes. ALBERT addresses issues such as memory limitations, longer training times, and unexpected model degration, and it aims to lightweight the model. RoBERTa uses a more dynamic masking strategy and uses Byte-Pair Encoding, and it demonstrates the best performance. Therefore, compared to the pre-trained baselines, including state-of-the-art models like KnowleNet and DynRT-Net, the proposed model demonstrates superior performance over other models.

Most of existing studies have conducted extensive experiments and comparisons on the MSD dataset. For a more comprehensive validation, we also test our proposed model on two additional publicly available dataset Memotion and MultiBully. For the fair comparison, we use the same evaluation metrics as other baselines. As shown in Tables 4 and 5, our proposed FSICN shows the best performance on Memotion and MultiBully. Compared to the state-of-the-art baseline, FSICN achieves an increase of 1.13% in F1-score on Memotion dataset, and achieves an increase of 2.57% in accuracy and 1.91% in F1-score on MultiBully dataset.

4.5. Ablation study

We conduct an ablation study to analyze the contribution of CMI, FIM, SIM and CFM modules of our proposed FSICN, as shown in Table 6. Firstly, we use only BERT and ViT as our baseline model (i.e., (a)) and fusion strategy is set to concatenation, the performance reaches its lowest point. When we introduce the CMI module on top of the baseline (i.e., (b)), performances achieve improvements of 1.19% in accuracy, 0.92% in binary-average F1-score, and 1.92% in macro-average F1-score. Compared to the concatenation structure that directly

Table 5
bluePerformance of FSICN compared to the baselines on MultiBully dataset with the evaluation metrics Acc, Precision, Recall and Binary-average F1-score.

Modality	Model	Acc	P	R	F1
Text	BERT-GRU	59.72	–	–	59.12
	RoBERTa	61.82	62.03	60.31	61.16
Image	ResNet	59.39	–	–	57.79
Image + Text	HFM	62.08	61.37	61.46	61.41
	CMGCN	62.51	61.88	62.14	62.01
	HKEmodel	62.75	62.43	62.61	62.52
	MILNET	63.44	62.58	62.19	62.38
	DIP	64.29	63.54	62.41	62.97
	CLIP	62.20	–	–	61.47
	KnowleNet	64.35	63.72	62.08	62.89
Ours	FSICN	**66.92**	**65.47**	**64.31**	**64.88**

Table 6
Ablation study of FSICN model on different components. **Base:** conducts multimodal sarcasm detection using only the backbone models (i.e., BERT and ViT). **CMI, FIM, SIM, CFM:** refer to the cross-modal interactive module, fact incongruity module, sentiment incongruity module and combination incongruity fusion module.

Setting	Components					MSD		
	Base	CMI	FIM	SIM	CFM	Acc↑	Binary-F1↑	Macro-F1↑
(a)	✓					85.23	83.47	84.16
(b)	✓	✓				86.42 (↑1.19)	84.39 (↑0.92)	86.08 (↑1.92)
(c)	✓	✓	✓			88.61 (↑3.38)	87.14 (↑3.67)	87.86 (↑3.70)
(d)	✓	✓		✓		87.33 (↑2.10)	86.55 (↑3.08)	86.92 (↑2.76)
(e)	✓	✓	✓	✓		89.66 (↑4.43)	88.35 (↑4.88)	89.42 (↑5.26)
(f)	✓	✓			✓	86.88 (↑1.65)	85.48 (↑2.01)	86.58 (↑2.42)
(g)	✓	✓	✓	✓	✓	**90.55 (↑5.32)**	**89.72 (↑6.25)**	**90.29 (↑6.13)**

concatenates text and image vectors, CMI may inadvertently truncate the complete information contained within the text and images individually. However, we use the pre-trained BERT and ViT model, where the [CLS] token is utilized to represent the entire sequence information, somewhat reducing information loss. Under this premise, cross-modal interaction way enables better recognition of the inherent relationship between text and image. This suggests that the cross-modal interaction between text and images facilitates information complementarity, leading to an improvement in sarcasm detection.

Then, concerning the fine-grained modules, we observe a further enhancement in the performance of all three metrics when adding the FIM (i.e., (c)) and SIM (i.e., (d)) modules, respectively. Upon the addition of both the FIM and SIM modules (i.e., (e)), the performances reach their peak, emphasizing the critical role of factual differences and sentiment clues within the FIM and SIM modules in recognizing sarcasm. Notably, using only the FIM module outperforms using the SIM model alone. Due to the subtlety of sarcasm, sentiment clues may be implicitly conveyed, making the FIM module more conducive to sarcasm detection. In addition, comparing (b) and (f) to (e) and (g) reveals that fusion strategy based on element-wise product outperform concatenation strategy. This highlights the superior effectiveness of non-linear fusion in enhancing modal representations. Finally, when we use all components, our proposed FSICN outperforms the baseline model by an increase of 5.32% in accuracy, 6.25% in binary-F1, and 6.13% in macro-F1. This effectiveness of our proposed model in multimodal sarcasm detection.

4.6. Influence of the number of DC and GCN layers

In this section, we examine how the number of DC layers in the fact incongruity module and the number of GCN layers in the sentiment incongruity module affect model performance, as illustrated in Figs. 3 and 4.

Initially, we observe a steady improvement as the number of dynamic connection layers increases from 1 to 3, reaching its peak at the 3rd layer. However, with further increases in the number of layers, performance gradually diminishes. We attribute this phenomenon to the model's gradual enhancement in learning capacity from 1 to 3 layers, allowing it to adaptively capture the most appropriate image-text pairs through multi-head dynamic routing layers. Nevertheless, as the number of layers continues to increase, the graph attention and masking mechanisms may introduce cross-modal relationships that are unrelated to the current image-text pair into the nodes, resulting in a performance decline.

Secondly, within the sentiment incongruity module, the model achieves its peak performance when the number of GCN layers is set to 2. Nevertheless, as the number of layers increases beyond this point, the model's performance gradually declined. We believe that this phenomenon is due to the occurrence of over-smoothing, which makes the features of all nodes increasingly similar. Hence, we set the number of DC layers to 3 and GCN layers to 2 for optimal performance.

4.7. Case study

To gain a deeper understanding of the principles underlying FIM and SIM in FSICN, we visualize the dynamic connection layer in FIM and the cross-modal graph matrix in EIM, as shown in Figs. 5 and 6. Firstly, from Fig. 5, we observe that tokens in the text fail to correspond well with the corresponding image patch in the first and second dynamic connection layers. This phenomenon indicates that the dynamic routing layers fail to learn the graph attention weights to make reasonable selections of image-text pairs. As the number of layers increases to 3, the multi-head dynamic routing layer calculates dynamic routing probability weights using the masked matrix to obtain the most suitable image-text pairs adaptively. Therefore, tokens can accurately focus on their corresponding image regions. For instance, when the dynamic connection layer is set to 3, "gorgeous" can concentrate well on the areas representing rain and windbell, while "day" focuses on the overcast region. This demonstrates that the designed FIM effectively discriminate the factual relationships between textual descriptions and image objects.

Secondly, Fig. 6 shows that the "gorgeous" exhibits higher weights in image regions ① representing overcast weather, image region ② representing rainfall, and image region ③ representing windbell. Among them, the weight is most pronounced in image region ②. The highly positive sentiment expressed by the word "gorgeous" conflicts with the gloomy colors associated with rainy weather, and the regions with larger weights represent highly correlated sentiment incongruity sarcasm clues across modalities. This validates the effectiveness of the proposed sentiment incongruity module in multi-modal sarcasm detection.

5. Conclusion

In this paper, we aim to explore multimodal sarcasm detection from a new perspective, introducing a fact-sentiment incongruity combination network to capture the fine-grained incongruities. First, we designed a fact incongruity module that contains dynamic connection layer to select the most suitable image-text pair to capture fact incongruity. Then, we generate the cross-modal graph by reconstructing edge weights to retrieve sentiment relations between text tokens and image objects to extract the sentiment incongruity. Furthermore, we construct a combination incongruity fusion layer to fuse the fact and sentiment incongruity, and introduce a cross-modal contrastive loss to further enhance the incongruity representations. Experiments and further analyses on the publicly available datasets demonstrate the improvements of our proposed model. Future works will focus on taken external large-scale language knowledge bases and large language modeling into account, which may result in its limited ability to effectively recognize metaphors. In addition, we would like to utilize Large Language Models to explore the affective relations of multimodal sarcasm detection in the future work.

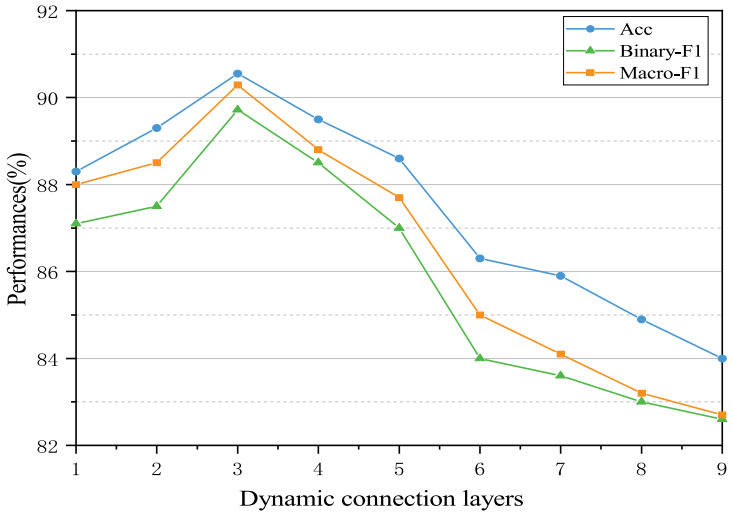

Fig. 3. Influence of the numbers of DC layers with the evaluation metrics.

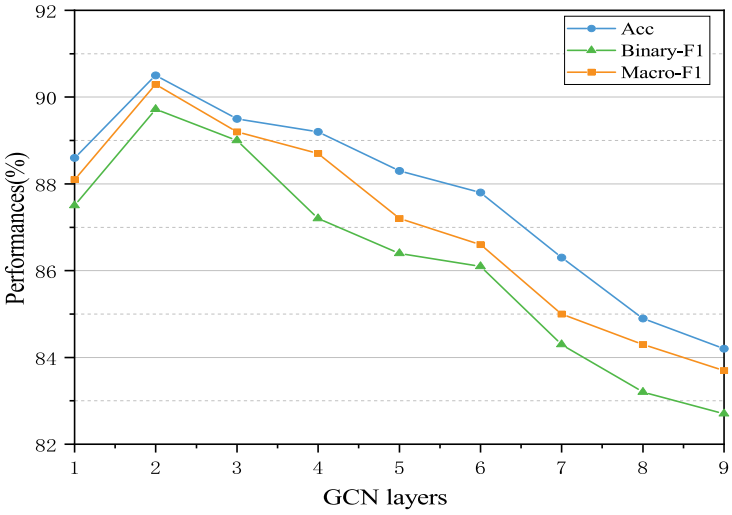

Fig. 4. Influence of the numbers of GCN layers with the evaluation metrics.

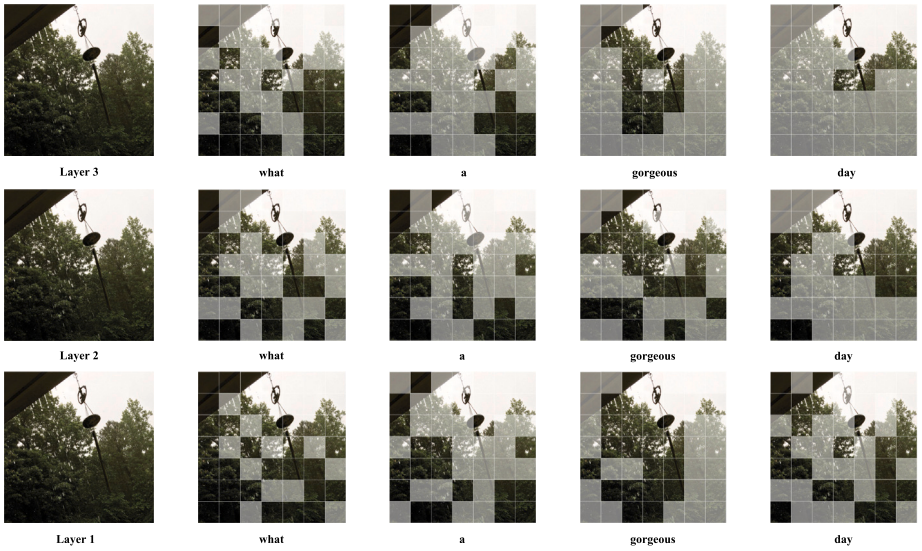

Fig. 5. Visualization of the dynamic connection layer.

Fig. 6. Visualization of the cross-modal graph weight matrix.

CRediT authorship contribution statement

Qiang Lu: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Yunfei Long:** Methodology, Writing – review & editing. **Xia Sun:** Funding acquisition, Methodology, Supervision, Writing – review & editing. **Jun Feng:** Investigation, Validation. **Hao Zhang:** Data curation, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant No. 61877050, Key R and D plan of Xianyang, China under Grant No. 2021ZDYF-GY-0033, and 2022 Yulin Science and Technology Plan Project, China under Grant No. CXY-2022-177

References

[1] S. Dews, E. Winner, Muting the meaning a social function of irony, Metaphor Symbol 10 (1) (1995) 3–19.
[2] R.W. Gibbs, On the psycholinguistics of sarcasm., J. Exp. Psychol. Gen. 115 (1) (1986) 3.
[3] R.W. Gibbs, On the psycholinguistics of sarcasm, Irony Lang. Thougt Cogn. Sci. Reader (2007) 173–200.
[4] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.
[5] S. Kumar, I. Mondal, M.S. Akhtar, T. Chakraborty, Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 11, 2023, pp. 12986–12994.
[6] Y. Liu, Y. Zhang, Q. Li, B. Wang, D. Song, What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability, in: Findings of the Association for Computational Linguistics, EMNLP 2021, 2021, pp. 871–880.
[7] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in twitter with hierarchical fusion model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2506–2515.
[8] N. Xu, Z. Zeng, W. Mao, 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, 2020, pp. 3777–3786.
[9] M. Bedi, S. Kumar, M.S. Akhtar, T. Chakraborty, Multi-modal sarcasm detection and humor classification in code-mixed conversations, IEEE Trans. Affect. Comput. (2021).
[10] D.S. Chauhan, G.V. Singh, A. Arora, A. Ekbal, P. Bhattacharyya, An emoji-aware multitask framework for multimodal sarcasm detection, Knowl.-Based Syst. 257 (2022) 109924.
[11] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, L. Nie, Mutual-enhanced incongruity learning network for multi-modal sarcasm detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 8, 2023, pp. 9507–9515.
[12] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, R. Xu, Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4707–4715.
[13] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, 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, 2022, pp. 1767–1777.
[14] T. Yue, R. Mao, H. Wang, Z. Hu, E. Cambria, KnowleNet: Knowledge fusion network for multimodal sarcasm detection, Inf. Fusion 100 (2023) 101921.
[15] H. Liu, W. Wang, H. Li, 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, 2022, pp. 4995–5006.
[16] C. Wen, G. Jia, J. Yang, DIP: Dual incongruity perceiving network for sarcasm detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2540–2550.
[17] Y. Li, H. Zhou, Y. Yin, J. Gao, Multi-label pattern image retrieval via attention mechanism driven graph convolutional network, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 300–308.
[18] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10921–10930.
[19] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, H.T. Shen, Cross-modal attention with semantic consistence for image–text matching, IEEE Trans. Neural Netw. Learn. Syst. 31 (12) (2020) 5412–5425.
[20] D. Sperber, D. Wilson, Précis of relevance: Communication and cognition, Behav. Brain Sci. 10 (4) (1987) 697–710.
[21] N. Babanejad, H. Davoudi, A. An, M. Papagelis, Affective and contextual embedding for sarcasm detection, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 225–243.
[22] D. Bamman, N. Smith, Contextualized sarcasm detection on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 9, No. 1, 2015, pp. 574–577.
[23] A. Joshi, P. Bhattacharyya, M.J. Carman, Automatic sarcasm detection: A survey, ACM Comput. Surv. 50 (5) (2017) 1–22.
[24] R. Schifanella, P. De Juan, J. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 1136–1145.
[25] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[26] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.
[27] H. Pan, Z. Lin, P. Fu, Y. Qi, W. Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: Findings of the Association for Computational Linguistics, EMNLP 2020, 2020, pp. 1383–1392.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[29] S. Pramanick, A. Roy, V.M. Patel, Multimodal learning using optimal transport for sarcasm and humor detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3930–3940.

[30] X. Wang, X. Sun, T. Yang, H. Wang, Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data, in: Proceedings of the First International Workshop on Natural Language Processing beyond Text, 2020, pp. 19–29.

[31] Z. Yin, F. You, Multimodal sarcasm semantic detection based on inter-modality incongruity, in: International Conference on Computer Graphics, Artificial Intelligence, and Data Processing, Vol. 12168, ICCAID 2021, SPIE, 2022, pp. 501–505.

[32] Z. Li, Q. Guo, Y. Pan, W. Ding, J. Yu, Y. Zhang, W. Liu, H. Chen, H. Wang, Y. Xie, Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis, Inf. Fusion (2023) 101891.

[33] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, R. Xu, Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 7124–7135.

[34] Q. Lu, X. Sun, Z. Gao, Y. Long, J. Feng, H. Zhang, Coordinated-joint translation fusion framework with sentiment-interactive graph convolutional networks for multimodal sentiment analysis, Inf. Process. Manage. 61 (1) (2024) 103538.

[35] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, T. Sun, Sentiment analysis: Comprehensive reviews, recent advances, and open challenges, IEEE Trans. Neural Netw. Learn. Syst. (2023).

[36] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M.A. Fadhel, M. Al-Amidie, L. Farhan, Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, J. Big Data 8 (2021) 1–74.

[37] Y. Long, R. Xiang, Q. Lu, C.R. Huang, M. Li, Improving attention model based on cognition grounded data for sentiment analysis, IEEE Trans. Affect. Comput. 12 (4) (2019) 900–912.

[38] J. Shen, M.D. Ma, R. Xiang, Q. Lu, E.P. Vallejos, G. Xu, C.R. Huang, Y. Long, Dual memory network model for sentiment analysis of review text, Knowl.-Based Syst. 188 (2020) 105004.

[39] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, et al., Pre-trained models: Past, present and future, AI Open 2 (2021) 225–250.

[40] Q. Lu, X. Sun, R. Sutcliffe, Y. Xing, H. Zhang, Sentiment interaction and multi-graph perception with graph convolutional networks for aspect-based sentiment analysis, Knowl.-Based Syst. 256 (2022) 109840.

[41] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th International Conference on Data Mining, ICDM, IEEE, 2016, pp. 439–448.

[42] M. Chen, S. Wang, P.P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, Multimodal sentiment analysis with word-level fusion and reinforcement learning, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 163–171.

[43] C. Zhu, M. Chen, S. Zhang, C. Sun, H. Liang, Y. Liu, J. Chen, SKEAFN: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis, Inf. Fusion 100 (2023) 101958.

[44] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, X. Huang, Sentiment-aware multimodal pre-training for multimodal sentiment analysis, Knowl.-Based Syst. 258 (2022) 110021.

[45] Z. Liu, B. Zhou, D. Chu, Y. Sun, L. Meng, Modality translation-based multimodal sentiment analysis under uncertain missing modalities, Inf. Fusion 101 (2024) 101973.

[46] J. Huang, Z. Lin, Z. Yang, W. Liu, Temporal graph convolutional network for multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 239–247.

[47] D. Wang, C. Tian, X. Liang, L. Zhao, L. He, Q. Wang, Dual-perspective fusion network for aspect-based multimodal sentiment analysis, IEEE Trans. Multimed. (2023).

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16×16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[49] Y. Tian, N. Xu, R. Zhang, W. Mao, 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), 2023, pp. 2468–2480.

[50] Y. Zhou, T. Ren, C. Zhu, X. Sun, J. Liu, X. Ding, M. Xu, R. Ji, Trar: Routing the attention spans in transformer for visual question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2074–2084.

[51] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al., Graph attention networks, stat 1050 (20) (2017) 10–48550.

[52] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, R. Xu, Affective dependency graph for sarcasm detection, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1844–1849.

[53] X. Yang, S. Feng, Y. Zhang, D. Wang, Multimodal sentiment detection based on multi-channel graph neural networks, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 328–339.

[54] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 105–114.

[55] C. Sharma, D. Bhageria, W. Scott, S. Pykl, A. Das, T. Chakraborty, V. Pulabaigari, B. Gambäck, SemEval-2020 task 8: Memotion analysis-the visuo-lingual metaphor! in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 759–773.

[56] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1739–1749.

[57] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.

[58] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1746–1751.

[59] Y. Tay, A.T. Luu, S.C. Hui, J. Su, Reasoning with sarcasm by reading in-between, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1010–1020.

[60] T. Xiong, P. Zhang, H. Zhu, Y. Yang, Sarcasm detection with self-matching networks and low-rank bilinear pooling, in: The World Wide Web Conference, 2019, pp. 2115–2124.

[61] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv preprint arXiv:1909.11942.

[62] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, Q.V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Adv. Neural Inf. Process. Syst. 32 (2019).

[63] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.

[64] G.-A. Vlad, G.-E. Zaharia, D.C. Cercel, C. Chiru, S. Trausan-Matu, UPB at SemEval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1208–1214.

[65] I. Chaturvedi, C.L. Su, R.E. Welsch, Fuzzy aggregated topology evolution for cognitive multi-tasks, Cogn. Comput. 13 (2021) 96–107.

[66] D.S. Chauhan, S. Dhanush, A. Ekbal, P. Bhattacharyya, Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4351–4360.

[67] R.A. Potamias, G. Siolas, A.G. Stafylopatis, A transformer-based approach to irony and sarcasm detection, Neural Comput. Appl. 32 (2020) 17309–17320.

[68] L.H. Li, M. Yatskar, D. Yin, C.J. Hsieh, K.W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019, arXiv preprint arXiv:1908.03557.

[69] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).