Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network

Bin Liang[1,2*], **Chenwei Lou**[1,2*], **Xiang Li**[1,2], **Min Yang**[3], **Lin Gui**[4], **Yulan He**[4,5], **Wenjie Pei**[1], **and Ruifeng Xu**[1,6†]

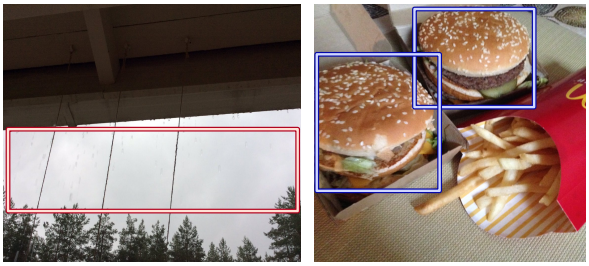
[1]School of Computer Science and Technology,
Harbin Institute of Technology, Shenzhen, China
[2] Joint Lab of HITSZ and China Merchants Securities, Shenzhen, China
[3] SIAT, Chinese Academy of Sciences, Shenzhen, China
[4] Department of Computer Science, University of Warwick, UK
[5] The Alan Turing Institute, UK, [6] Peng Cheng Laboratory, Shenzhen, China
{bin.liang, xiangli}@stu.hit.edu.cn, louchenw@163.com
min.yang@siat.ac.cn, {lin.gui, Yulan.He}@warwick.ac.uk
wenjiecoder@outlook.com, xuruifeng@hit.edu.cn

Abstract

With the increasing popularity of posting multimodal messages online, many recent studies have been carried out utilizing both textual and visual information for multi-modal sarcasm detection. In this paper, we investigate multi-modal sarcasm detection from a novel perspective by constructing a cross-modal graph for each instance to explicitly draw the ironic relations between textual and visual modalities. Specifically, we first detect the objects paired with descriptions of the image modality, enabling the learning of important visual information. Then, the descriptions of the objects are served as a bridge to determine the importance of the association between the objects of image modality and the contextual words of text modality, so as to build a cross-modal graph for each multi-modal instance. Furthermore, we devise a cross-modal graph convolutional network to make sense of the incongruity relations between modalities for multi-modal sarcasm detection. Extensive experimental results and in-depth analysis show that our model achieves state-of-the-art performance in multi-modal sarcasm detection[1].

1 Introduction

Sarcasm is a peculiar form of sentiment expressions, allowing individuals to express contempt sentiment or intention that is converse to the authentic/apparent sentiment information (Gibbs, 1986; Dews and Winner, 1995; Gibbs, 2007). As such, accurately detecting satirical/ironic expression could potentially improve the performance of sentiment analysis and opinion mining (Pang and Lee, 2008; Kumar Jena et al., 2020; Pan et al., 2020).



(a) What a wonderful weather! (b) Feeding my abs nothing but the best quality beef.

Figure 1: Two multi-modal sarcastic examples. Boxes and words in the same color denote highly correlated sarcastic cues.

In today's fast growing social media platforms, it is common to post multi-modal messages. Therefore, in addition to developing sarcasm detection models for textual data (Riloff et al., 2013; Joshi et al., 2015), it is increasingly popular to explore sarcasm detection in multi-modal data such as text and images (Schifanella et al., 2016; Cai et al., 2019). Dealing with multimodal data requires an understanding of the information presented in different modalities. As the sarcastic example shown in Figure 1 (a), text-only approaches may erroneously identify it as a positive sentiment expression due to the phrase "*wonderful weather*". This post however contains a sarcastic expression with negative sentiment, because it is accompanied by an image with "*thunderstorm clouds*". The key of effective multi-modal sarcasm detection is to accurately extract the incongruent sentiment cues from

[*] The first two authors contribute equally to this work.
[†] Corresponding Author.
[1]The source code of this work is released at https://github.com/HITSZ-HLT/CMGCN.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1767 - 1777

different modalities, allowing the detection of the true sentiment conveyed in the message.

To perform multi-modal sarcasm detection on data composed of text and image, several related research efforts attempt to concatenate the textual and visual features to fuse sarcastic information (Schifanella et al., 2016), employ attention mechanism to implicitly fuse the features of different modalities based on external knowledge (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020), or build interactive graphs to model the relations of different modalities (Liang et al., 2021a). Despite promising progress made by existing models, they still suffer from the following limitations: **1)** Simply considering the whole image does not produce good results, mostly due to the intricate visual information presented in an image; not to mention that only particular visual patches are related to the text. As in the examples shown in Figure 1, the correct results can be easily obtained by only tracking the visual information in the bounding boxes. Therefore, discriminating key visual objects from the irrelevant ones could lead to improved learning of visual information. **2)** Crucial visual information that relates to the sarcastic cues of text modality may be scattered in an image (Figure 1 (b)). As such, it is essential to focus on drawing the intricate sentiment connections between text and image modalities, allowing a good exploitation of the contradictory sentiment information between modalities for learning sarcastic clues.

To this end, we propose a novel cross-modal graph convolutional networks (CMGCN) by constructing a cross-modal graph for each instance, where the important visual information and the related textual tokens are explicitly linked. This allows for the extraction of incongruous implications between two modalities in sarcasm detection. Concretely, instead of trying to produce a caption of the whole image, we first detect the objects of the image to capture the important visual regions and the corresponding *attribute-object* pairs via the approach proposed by Anderson et al. (2018). Then, we explore a novel solution to assign weights to the edges of the cross-modal graph by means of computing the word similarities between the *object* descriptors of the *attribute-object* pairs and textual words based on the WordNet (Miller, 1992). Further, to introduce the multi-modal sentiment relations into the cross-modal graphs, inspired by (Lou et al., 2021), we devise a modulating factor of sentiment relation for each edge by retrieving the affective weights of *attribute* descriptors (usually adjectives with affective information) and textual words from external affective knowledge (SenticNet (Cambria et al., 2020)). As such, the modulating factors can be adopted to refine the edge weights of word similarities, allowing the capture of sentiment incongruities of the cross-modal nodes in the graph. Further, in the light of cross-modal graphs, we deploy a GCN architecture to make sense of the incongruous relations across the modalities for multi-modal sarcasm detection.

The main contributions of our work are summarized as follows:

- To the best of our knowledge, we are the first to explore the use of the graph model based on auxiliary object detection for modeling the contradictory sentiments between key textual and visual information in multi-modal sarcasm detection.
- Using the attribute-object pairs of the image objects as the bridge, a novel approach of constructing cross-modal graphs is developed to explicitly link the two modalities by edges with the varying degree of importance.
- A series of experiments on a publicly available multi-modal sarcasm detection benchmark dataset show that our proposed method achieves the state-of-the-art performance.

2 Related Work

2.1 Multi-modal Sarcasm Detection

Previous work of sarcasm detection has been applied to textual utterances information (Zhang et al., 2016; Tay et al., 2018; Babanejad et al., 2020). Different from text-based sarcasm detection, multi-modal sarcasm detection aims to identify the sarcastic expression among different modalities (Schifanella et al., 2016; Castro et al., 2019). Schifanella et al. (2016) firstly tackled the multi-modal sarcasm detection task with text and image modalities by manually designed features. Cai et al. (2019) created a new dataset and proposed a hierarchical fusion model for multi-modal sarcasm detection. Xu et al. (2020) explored decomposition and relation network to model both cross-modality contrast and semantic association in sarcasm detection. Pan et al. (2020) proposed inter-modality attention and co-attention to learn the contradiction of sarcasm. For

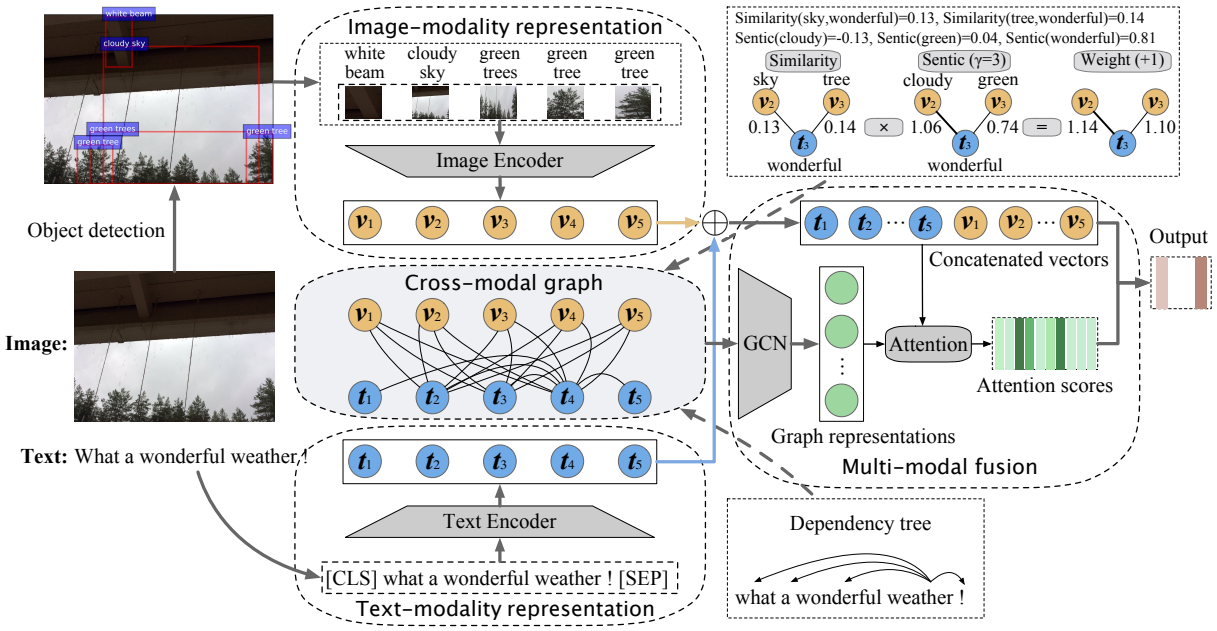


Figure 2: The architecture of the proposed CMGCN. $\oplus$ represents matrix concatenation.

the graph-based methods, Liang et al. (2021a) deployed a heterogeneous graph structure to learn the sarcastic features from both intra- and inter-modality perspectives. However, this method tried to grasp the visual information of the whole image, and meanwhile ignore the sentiment expression between different modalities. Therefore, different from (Liang et al., 2021a), we explore a novel cross-modal GCN model based on the important visual information and sentiment cues to leverage the inconsistent implications between different modalities and thus improve the performance of multi-modal sarcasm detection.

2.2 Graph Neural Networks

Models based on graph neural networks (GNN), including graph convolutional network (GCN) (Kipf and Welling, 2017) and graph attention network (GAT) (Velickovic et al., 2018), have achieved promising performance in many recent research studies, such as visual representation learning (Wu et al., 2019; Xie et al., 2021), text representation learning (Yao et al., 2019; Lou et al., 2021; Liang et al., 2021b, 2022), and recommendation systems (Ying et al., 2018; Tan et al., 2020). Further, there are also some research studies explored graph models to deal with the multi-modal tasks, such as multi-modal sentiment detection (Yang et al., 2021), multi-modal named entity recognition (Zhang et al., 2021), cross-modal video moment retrieval (Zeng et al., 2021), multi-modal neural machine translation (Yin et al., 2020), and multi-modal sarcasm detection (Liang et al., 2021a).

3 Methodology

In this section, we describe our proposed Cross-Modal Graph Convolutional Networks (**CMGCN**) model for multi-modal sarcasm detection in details. As demonstrated in Figure 2, the architecture of the proposed CMGCN contains four main components: 1) *Text-modality representation*, which employs the pre-trained uncased BERT-base model (Devlin et al., 2019) as the text encoder to capture the hidden representation of the text-modality; 2) *Image-modality representation*, which deploys the pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021) as the image encoder to capture the hidden representation of the image-modality with respect to each bounding box (visual region); 3) *Cross-modal graph*, which constructs a cross-modal graph for each multi-modal example based on the external affective knowledge source and the hidden representations of text and image modalities; 4) *Multi-modal fusion*, which fuses the representations from image and text modalities to capture the sarcastic features by means of a GCN structure and an attention mechanism.

3.1 Text-modality Representation

For text processing, given a sequence of words $s = \{w_i\}_{i=1}^{n}$, n is the length of the text s. We first adopt the pre-trained uncased BERT-base

model (Devlin et al., 2019) to map each word w_i into a d^T-dimensional embedding:

$$\boldsymbol{X}^T = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] = \text{BERT}([\text{CLS}]s[\text{SEP}]) \tag{1}$$

Where $\boldsymbol{X}^T$ is the embedding matrix of the input text. Here, the representations of tokens `[CLS]` and `[SEP]` are not utilized in constructing the cross-modal graph. Subsequently, to unify the dimensions of representations between different modalities and capture the sequential relations of the context, we utilize a bidirectional LSTM (Bi-LSTM) to learn the text-modality representation of the input text:

$$\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_n\} = \text{Bi–LSTM}(\boldsymbol{X}^T) \tag{2}$$

Where $\boldsymbol{t}_j \in \mathbb{R}^{2d_h}$ denotes the hidden state vector at time step j from the bidirectional LSTM, d_h denotes the dimensionality of the text-modality hidden state representation.

3.2 Image-modality Representation

For image processing, given an image I, we first adopt a trained toolkit proposed by Anderson et al. (2018) to derive a series of bounding boxes (objects) paired with their attribute-object pairs. For each visual region of the bounding box $I_i \in \mathbb{R}^{L_h \times L_w}$, following (Xu et al., 2020), we first resize it to 224×224, i.e. $L = L_h = L_w = 224$. Subsequently, following (Dosovitskiy et al., 2021), we reshape the region $I_i \in \mathbb{R}^{L \times L}$ into a sequence $I_i = \{\boldsymbol{p}_j \in \mathbb{R}^{L/p \times L/p}\}_{j=1}^{r}$, where $r = p \times p$ is the number of patches. Then, we flatten and map each patch to a d^I-dimensional vector with a trainable linear projection: $\boldsymbol{z}_j = \boldsymbol{p}_j \boldsymbol{E}$.

For each sequence of image patches, a `[class]` token embedding $\boldsymbol{z}_{[class]} \in \mathbb{R}^{d^I}$ is prepended for the sequence of embedded patches, and position embeddings are added to the patch embeddings to retain positional information. The input of each visual region I_i is represented as:

$$\boldsymbol{Z}_i = [\boldsymbol{z}_{[class]}; \boldsymbol{z}_1; \boldsymbol{z}_2; \cdots; \boldsymbol{z}_r] + \boldsymbol{E}_{pos} \tag{3}$$

Where $\boldsymbol{Z}_i \in \mathbb{R}^{(r+1) \times d^I}$ is the input matrix of the image patches, and $\boldsymbol{E}_{pos} \in \mathbb{R}^{(r+1) \times d^I}$ is the position embedding matrix. Then, we feed the input matrix $\boldsymbol{Z}_i$ into the ViT encoder to acquire the representation $\boldsymbol{h}_i$ of visual region I_i:

$$\boldsymbol{H}_i = \text{ViT}(\boldsymbol{Z}_i),\ \boldsymbol{h}_i = \boldsymbol{H}_{i,[class]} \tag{4}$$

We use the representation of the `[class]` token embedding to represent the visual region. Finally, the representation of the image I is defined as:

$$\boldsymbol{X}^I = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_m\} \tag{5}$$

Where m is the number of visual regions.

Subsequently, we employ a trainable Linear Projection to map each $\boldsymbol{v}_i$ to a $2d_h$-dimensional vector:

$$\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\} = \boldsymbol{X}^I \boldsymbol{W}^V \tag{6}$$

Where $\boldsymbol{W}^V \in \mathbb{R}^{d^I \times 2d_h}$ is a trainable parameter.

3.3 Cross-modal Graph

In this section, we describe how to construct a cross-modal graph. To leverage the relations between multi-modal features, we employ a graph structure to link the textual words with the associated image objects. Here, the nodes of the cross-modal graph are the representations of text and image modalities. Many GCN-based approaches have demonstrated that the weights of the edges are crucial in graph information aggregation (Liang et al., 2021b; Yang et al., 2021; Lou et al., 2021). As such, constructing a cross-modal graph boils down to the setting of the edge weights in the graph.

To this end, we explore a novel approach of setting the weights based on both word similarities and affective clues between textual words and the *attribute-object* pairs of the image regions, and the dependency tree of the text-modality. The adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{(n+m) \times (n+m)}$ of the cross-modal graph is defined as:

$$A_{i,j} = \begin{cases} 1 & \text{if } \mathcal{D}_{i,j} \text{ and } i < n, j < n \\ \boldsymbol{\kappa}_{i,j} & \text{if } i < n, j \geq n \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$\boldsymbol{\kappa}_{i,j} = Sim(w_i, o_j) \times \boldsymbol{\xi}_{i,j} + 1 \tag{8}$$

$$\boldsymbol{\xi}_{i,j} = \gamma^{-\boldsymbol{\omega}(w_i)\boldsymbol{\omega}(a_j)} \times |\boldsymbol{\omega}(w_i) - \boldsymbol{\omega}(a_j)| \tag{9}$$

Where $\mathcal{D}_{i,j}$ indicates that there is a relation between w_i and w_j in the dependency tree of the sentence. $Sim(\cdot)$ represents the computation of word similarity[2]. We set $Sim(\cdot) = 0$ if the return value is $None$. $\boldsymbol{\xi}_{i,j}$ is a modulating factor refers to the sentiment relation (sentiment incongruity) between an image region and a text token. $\boldsymbol{\omega}(w_i) \in [-1, 1]$ represents the affective weight of

[2]We employ the NLTK toolkit (`http://www.nltk.org/`) to compute the similarity of a word pair based on the WordNet.

word w_i retrieved from SenticNet (Cambria et al., 2020). We set $\boldsymbol{\omega}(w_i) = 0$ if w_i cannot be found in SenticNet. $|\cdot|$ represents absolute value calculation. a_j and o_j respectively denote the *attribute* and the *object* of the bounding box j. Inspired by Kipf and Welling (2017), we construct the cross-modal graph as an undirected graph, $A_{i,j} = A_{j,i}$, and set a self-loop for each node, $A_{i,i} = 1$.

The intention of the cross-modal graph construction (Equations 7 and 9) is that: 1) As in the examples shown in Figure 1, the sarcastic information of text-modality may be expressed by multiple words, such as "*wonderful weather*". Therefore, we incorporate the syntax-aware relations over the dependency tree of the sentence into the cross-modal graph to advance the learning of the contextual dependencies[3]. 2) We devise a coefficient $\boldsymbol{\kappa}_{i,j}$, which is associated with the affective weights, to modulate the influence of contrary sentiment relations. Here, $\gamma > 1$ is a tuned hyper-parameter to regulate the bias of inconsistent sentiment relations. That is, if the polarities of $\boldsymbol{\omega}(w_i)$ and $\boldsymbol{\omega}(a_j)$ are opposite, the value of γ is boosted, otherwise the value is shrunk. Especially, the greater the affective weights, the higher the confidence that the value of γ is boosted or shrunk. 3) We add 1 to the cross-modal edges to pay more attention to the cross-modal nodes aggregation.

3.4 Multi-modal Fusion

For each instance, we explore a graph architecture to extract the crucial sarcastic clues by aggregating the correlation of nodes in the cross-modal graph. Concretely, we feed the adjacency matrix of the cross-modal graph $\boldsymbol{A}$ and the corresponding nodes' representations $\boldsymbol{R}$ of each multi-modal example into a multi-layers GCNs architecture to derive the graph representation. For each graph convolutional operation, each node in the l-th GCN layer is updated according to the hidden representations of its neighborhoods according to the adjacency matrices of the cross-modal graph, which is defined as:

$$\boldsymbol{G}_l = \mathrm{ReLU}(\tilde{\boldsymbol{A}}\boldsymbol{G}_{l-1}\boldsymbol{W}_l + \boldsymbol{b}_l) \tag{10}$$

Where $\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix. $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}$, where $D_{ii} = \sum_j A_{i,j}$. $\boldsymbol{G}_{l-1}$ is the hidden graph representation evolved from the preceding GCN layer. $\boldsymbol{W}_l \in \mathbb{R}^{2d_h \times 2d_h}$, $\boldsymbol{b}_l \in \mathbb{R}^{2d_h}$ are the trainable parameters of the l-th GCN layer. The nodes input of the first GCN layer are the concatenation of text-modality and image-modality representations: $\boldsymbol{G}_0 = \boldsymbol{R}$. Here, $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{n+m}\} = \{\boldsymbol{t}_1, \cdots, \boldsymbol{t}_n, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_m\}$.

Subsequently, inspired by (Zhang et al., 2019), we employ a retrieval-based attention mechanism to capture the graph-oriented attention information from the concatenation of text and image representations $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{n+m}\}$ by means of the graph representation $\boldsymbol{g}$ derived from the final GCN layer. The intention is to retrieve crucially associated cross-modal features that are explicitly connected in the cross-modal graph. The attention weights are computed as:

$$\alpha_t = \frac{\exp(\beta_t)}{\sum_{i=1}^{n+m}\exp(\beta_i)}, \quad \beta_t = \sum_{i \in \mathcal{C}} \boldsymbol{r}_t^\top \boldsymbol{g}_i \tag{11}$$

Where $\mathcal{C}$ denotes a set of indices in which nodes contain cross-modal edges in the graph. $\top$ represents the matrix transposition. The final sarcastic representation is defined as:

$$\boldsymbol{f} = \sum_{t=1}^{n+m} \alpha_t \boldsymbol{r}_t \tag{12}$$

Then, the final sarcastic representation is fed into a fully-connected layer with a softmax function to capture a probability distribution $\hat{\boldsymbol{y}} \in \mathbb{R}^{d_p}$ in the sarcasm decision space:

$$\hat{\boldsymbol{y}} = \mathrm{softmax}(\boldsymbol{W}_o\boldsymbol{f} + \boldsymbol{b}_o) \tag{13}$$

Where d_p is the dimensionality of sarcasm labels. $\boldsymbol{W}_o \in \mathbb{R}^{d_p \times 2d_h}$ and $\boldsymbol{b}_o \in \mathbb{R}^{d_p}$ are trainable parameters.

3.5 Learning Objective

We minimize the cross-entropy loss via the standard gradient descent algorithm to train the model:

$$\min_{\Theta}\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{d_p} y_i^j \log\hat{y}_i^j + \lambda||\Theta||^2 \tag{14}$$

where N is the training data size. $\boldsymbol{y}_i$ and $\hat{\boldsymbol{y}}_i$ respectively represent the ground-truth and estimated label distribution of instance i. Θ denotes all trainable parameters of the model, λ represents the coefficient of L_2-regularization.

[3]We employ the spaCy toolkit (https://spacy.io/) to derive the dependency tree of a sentence.

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

Table 1: Statistics of the experimental data.

4 Experimental Setup

4.1 Dataset

We conduct experiments on a publicly available multi-modal sarcasm detection benchmark dataset collected by Cai et al. (2019). This dataset contains English tweets expressing *sarcasm* as *Positive* examples and those expressing *non-sarcasm* as *Negative* examples. Each example in the dataset consists of a text and an associated image. The statistics of the dataset are shown in Table 1.

4.2 Experimental Settings

For a fair comparison, the data preprocessing follows (Cai et al., 2019). We set the maximum number of visual regions as 10 for object detection results. That is, we select the top 10 bounding boxes with highest scores if the objects are greater than 10. We utilize the pre-trained uncased BERT-base (Devlin et al., 2019) module to embed each word of text-modality as a 768-dimensional embedding and employ the pre-trained ViT[4] (Dosovitskiy et al., 2021) to embed each visual region patch as a 768-dimensional embedding, i.e. $d^T = d^I = 768$. The resolution of visual region patch is set to $L_p = 32$, correspondingly, $p = 7, r = 49$.[5] The number of GCN layers is set to 2, which is the optimal depth in the pilot experiments. The dimensionality of hidden representations is set to $d_h = 512$. The coefficient λ is set to 0.00001. Adam is utilized as the optimizer with a learning rate of 0.00002, and the mini-batch size is 32. The dropout rate with 0.1 is utilized to avoid overfitting. We use early-stopping with patience of 5. We set $\gamma = 3$ to compute the modulating factor of incongruous multi-modal sentiment relations, which is the optimal hyper-parameter in the pilot experiments.

Following (Cai et al., 2019), we use *Accuracy*, *Precision*, *Recall*, and *F1-score* to measure the model performance. Since the label distribution of the dataset is imbalanced, following (Pan et al., 2020), we also report Macro-average results. The experimental results of our models are averaged over 10 runs with different random seeds to ensure the final reported results are statistically stable.

[4] https://github.com/lukemelas/PyTorch-Pretrained-ViT

[5] We also tried other division resolutions, and found that the fluctuation of performance is negligible over different resolutions of image patches.

4.3 Comparison Models

We compare our proposed **CMGCN** model with a series of strong baselines, summarized as follow:

1) Image-modality methods: These models use only visual information for sarcasm detection, including **Image** (Cai et al., 2019), which employs ResNet (He et al., 2016) to train a sarcasm classifier; and **ViT** (Dosovitskiy et al., 2021), which utilizes the '`[class]`' token representation of the pre-trained ViT to detect the sarcasm.

2) Text-modality methods: These models use only textual information, including **TextCNN** (Kim, 2014), a deep learning model based on CNN for text classification; **Bi-LSTM**, a bidirectional LSTM network for text classification; **SIARN** (Tay et al., 2018), adopting inner-attention for textual sarcasm detection; **SMSD** (Xiong et al., 2019), exploring a self-matching network to capture textual incongruity information; and **BERT** (Devlin et al., 2019), the vanilla pre-trained uncased BERT-base taking '`[CLS] text [SEP]`' as input.

3) Multi-modal methods: These models take both text- and image-modality information. Including **HFM** (Cai et al., 2019), a hierarchical multimodal features fusion model for multi-modal sarcasm detection; **D&R Net** (Xu et al., 2020), a Decomposition and Relation Network modeling both cross-modality contrast and semantic association; **Res-BERT** (Pan et al., 2020), concatenating image features and BERT-based text features for sarcasm prediction; **Att-BERT** (Pan et al., 2020), exploring an inter-modality attention and a co-attention to model the incongruity of multi-modal sarcasm detection; and **InCrossMGs** (Liang et al., 2021a), a graph-based model to leverage the sarcastic relations from both intra- and inter-modal perspectives.

We also explore several variants of **CMGCN** to analyze the impact of different components in the ablation study: **1) w/o** $\mathcal{G}$ denotes without cross-modal graph, which only concatenates the representations of '`[class]`' and '`[CLS]`' tokens from ViT and BERT for sarcasm detection; **2) w/o** $\mathcal{O}$ denotes without object detection. The whole image is input into the image encoder, and the edge weights are set to 1 in the cross-modal graphs; **3) w/o** $\mathcal{S}$ denotes without using external knowledge.

Modality	Method	*Acc (%)*	*F1-score*			*Macro-average*		
			Pre (%)	*Rec (%)*	*F1 (%)*	*Pre (%)*	*Rec (%)*	*F1 (%)*
image	Image (Cai et al., 2019)	64.76♮	54.41♮	70.80♮	61.53♮	60.12	73.08	65.97
	ViT (Dosovitskiy et al., 2021)	67.83	57.93	70.07	63.43	65.68	71.35	68.40
text	TextCNN (Kim, 2014)	80.03♮	74.29♮	76.39♮	75.32♮	78.03	78.28	78.15
	Bi-LSTM	81.90♮	76.66♮	78.42♮	77.53♮	80.97	80.13	80.55
	SIARN (Tay et al., 2018)	80.57♮	75.55♮	75.70♮	75.63♮	80.34	78.81	79.57
	SMSD (Xiong et al., 2019)	80.90♮	76.46♮	75.18♮	75.82♮	80.87	78.20	79.51
	BERT (Devlin et al., 2019)	83.85♮	78.72♮	82.27♮	80.22♮	81.31	80.87	81.09
image+text	HFM (Cai et al., 2019)	83.44♮	76.57♮	84.15♮	80.18♮	79.40	82.45	80.90
	D&R Net (Xu et al., 2020)	84.02♮	77.97♮	83.42♮	80.60♮	-	-	-
	Res-BERT (Pan et al., 2020)	84.80♮	77.80	84.15	80.85	78.87♮	84.46♮	81.57♮
	Att-BERT (Pan et al., 2020)	86.05♮	78.63	83.31	80.90	80.87♮	85.08♮	82.92♮
	InCrossMGs (Liang et al., 2021a)	86.10♮	81.38♮	84.36♮	82.84♮	85.39♮	85.80♮	85.60♮
	CMGCN (ours)	**87.55**⋆	**83.63**⋆	**84.69**	**84.16**⋆	**87.02**⋆	**86.97**⋆	**87.00**⋆

Table 2: Main experimental results regarding unimodal and multimodal scenarios. The results of baselines with ♮ are retrieved from (Liang et al., 2021a), others are run by the open source codes. Best scores of each group are in bold. Results with ⋆ denote the significance tests of our CMGCN over the baseline models at $p-\text{value} < 0.05$.

All weights of edges are set to 1 in the cross-modal graph. Further, **4) w/o** $\mathcal{S}^w$ represents without using affective knowledge; **5) w/o** $\mathcal{D}$ denotes without using syntax-aware information of text-modality in graph construction.

Further, to investigate the effectiveness of our **CMGCN** when used with different pre-trained models, we also set the following variants:

1) -GloVe+ResNet: We replace BERT with GloVe (Pennington et al., 2014) to initialize each word into a 300-dimensional embedding and ViT with ResNet-152 (He et al., 2016) to embed each image patch as a 2048-dimensional vector.

2) -GloVe+ViT: We use GloVe as text encoder and use ViT as image encoder.

3) -BERT+ResNet: We use BERT as text encoder and use ResNet-152 as image encoder.

5 Experimental Results

5.1 Main Results

We report the comparison results regarding *Text-modality*, *Image-modality*, and *Text+Image modalities* in Table 2. From the results, we can draw the following conclusions. **1)** Our proposed **CMGCN** outperforms existing baselines across all metrics. This verifies the effectiveness of our proposed model in multi-modal sarcasm detection. **2)** We conduct significance tests of our **CMGCN** over the baseline models, the results show that our **CMGCN** significantly outperforms the baseline models in terms of most of the evaluation metrics (with $p-\text{value} < 0.05$). **3)** Our CMGCN model performs consistently better than the previous graph-based method (InCrossMGs), which demonstrates that recognizing significant visual regions and modeling sentiment relations can lead to improved performance. **4)** The methods based on text modality achieve consistently better performance than the methods based on image modality, which shows that the expression of sarcastic/non-sarcastic information primarily resides in the text modality. **5)** Methods based on both image and text modalities perform better than the unimodal baselines overall. This implies that leveraging the information of both image and text modalities is more effective for multi-modal sarcasm detection. **6)** The results of macro metrics are better than other commonly used metrics overall, which indicates that models perform better in the "*negative*" class due to the imbalanced class distribution.

Model	*Acc. (%)*	*F1 (%)*	*Macro-F1 (%)*
CMGCN	**87.55**	**84.16**	**87.00**
w/o $\mathcal{G}$	84.12	80.64	81.47
w/o $\mathcal{O}$	84.55	81.09	82.31
w/o $\mathcal{S}$	85.63	81.82	83.28
w/o $\mathcal{S}^w$	86.54	82.73	84.76
w/o $\mathcal{D}$	87.25	83.64	86.13

Table 3: Experimental results of ablation study.

5.2 Ablation Study

To analyze the impact of different components of our proposed **CMGCN**, we conduct an ablation study and report the results in Table 3. Note that removal of cross-modal graph (w/o $\mathcal{G}$) sharply degrades the performance, which verifies the significance of cross-modal in multi-modal features fusion for learning sarcastic expressions in multi-modal sarcasm detection. Removal of object de-

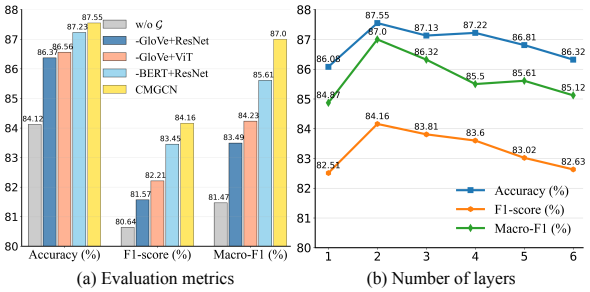


Figure 3: Performance of using different pre-trained methods (a) and using different GCN layers (b).

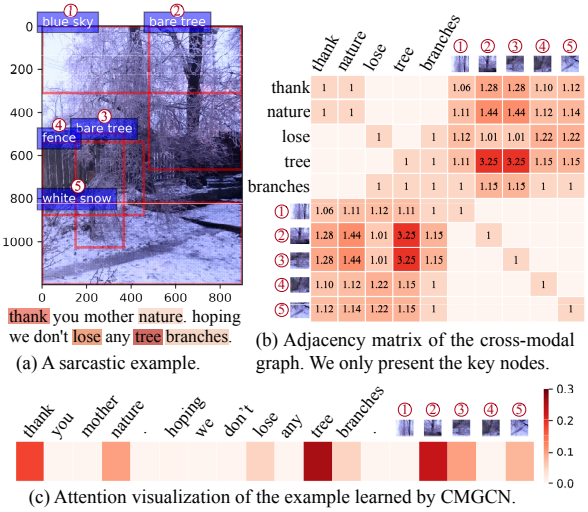


Figure 4: Visualization of a typical example.

tection (w/o $\mathcal{O}$) leads to considerable performance degradation, which demonstrates that adopting object detection to track important visual information is effective for constructing crucial relations between visual and textual information in the cross-modal graphs. From the results of w/o $\mathcal{S}$ and $\mathcal{S}^w$, we conclude that exploiting the *attribute-object* pair as a bridge to set edge weights based on word similarity is effective when constructing cross-modal graphs. Further, leveraging affective clues to capture multi-modal sentiment incongruity between text- and image-modality is effective in sarcasm detection, and thus leads to improved performance. In addition, removal of syntax-aware information of text-modality leads to slight performance degradation, which indicates that incorporating syntactic information in the graph makes better learning of dependency relations of textual words and thus improves the performance of sarcasm detection.

5.3 Generalizability of Cross-modal Graph

To investigate the generalizability and effectiveness of our proposed cross-modal graph when used with different pre-trained methods, we conduct experiments with five variants of our proposed **CMGCN** by using different text and image encoders. The experimental results are shown in Figure 3 (a). Note that the proposed cross-modal graph can directly work with various pre-trained models and performs consistently better than that without cross-modal graph (w/o $\mathcal{G}$). This demonstrates the generalizability and effectiveness of our proposed cross-modal graph in multi-modal sarcasm detection. Further, from the results, we can also conclude that superior performance is obtained when using more powerful pre-trained methods, such as ViT and BERT.

5.4 Impact of GCN Layers

In this section, we analyze the impact of the number of GCN layers on the performance of our proposed **CMGCN**. We vary the layer number from 1 to 6 and report the results in Figure 3 (b). Note that the 2-layer GCN architecture performs better than others overall, and thus the number of GCN layers is set to 2 in our model. Model with one layer performs worse, which indicates that a shallow graph network structure is not able to learn sarcastic features well. When the number of layers is greater than 2, the performance tends to decline. This shows that further increasing the number of layers beyond 2 degrades the model performance possibly due to the sharp increase of parameters.

5.5 Visualization

To qualitatively investigate how the proposed **CMGCN** works in multi-modal sarcasm detection, we present a visualization of cross-modal graph construction and attention values of a multi-modal sarcasm example. The results are shown in Figure 4. We first show a sarcasm example and its corresponding object detection results in Figure 4 (a). Note that the correct label of this example can be easily inferred if the relations of crucial sarcastic clues of text (marked by the light red color) and the corresponding visual regions are captured by the model. To demonstrate how the proposed **CMGCN** identifies the important sarcastic clues, we show the adjacency matrix of the cross-modal graph of this example in Figure 4 (b). Note that highly correlated sarcastic clues in different modalities are connected by edges with large weights in the graph. This verifies the effectiveness of the proposed cross-modal graph in learning multi-modal sarcastic information. Further, based on the cross-

modal graph, we show the attention visualization of this example in Figure 4 (c). The crucial textual tokens and the related image regions are highly attended by our proposed **CMGCN**, which helps identify the incongruity among the learned important features for learning sarcastic expressions and thus leads to improved performance of multi-modal sarcasm detection.

6 Conclusion and Future Work

This paper has proposed a novel cross-modal graph architecture for multi-modal sarcasm detection, in which the crucial visual regions can be explicitly connected to the highly correlated textual tokens for learning the incongruity sentiment of sarcastic expression. Specifically, unlike previous research efforts that simply consider the visual information of the whole image, we attempt to recognize the important visual regions via object detection results, and further devise a novel cross-modal graph to explicitly establish the connections of scattered visual regions and the associated textual tokens. More concretely, owing to the object detection results, the *attribute-object* pair descriptors of the objects are served as a bridge to track the highly related sarcastic cues between image and text modalities and their connection weights, and then deriving the cross-modal graphs based on external knowledge bases. Afterwards, a GCNs architecture based on a retrieval-based attention mechanism is employed to capture the key incongruity sentiment expressions across different modalities for multi-modal sarcasm detection. To the best of our knowledge, it is the first study of utilizing a cross-modal graph to extract intricate multi-modal sarcastic relations via object detection and sentiment cues from external knowledge bases. Extensive experiments on a public benchmark dataset show that our proposed approach significantly outperforms state-of-the-art baseline methods.

As described in Section 3.3, the weights of edges in the cross-modal graph are computed based on both word similarities and affective clues between textual words and the *attribute-object* pairs of the image regions, and the dependency tree of the text-modality. The approach can be easily generalized to other sentiment-related multi-modal learning scenarios. Nevertheless, the cross-graph solution might not be generalized well to other multi-modal tasks or data genres, if there is a lack of affective knowledge or a difficulty in deriving dependency trees in low-resource settings. Therefore, future research can consider exploiting alternatively approaches to automatically learn the weights of edges in the cross-modal graph without relying on external knowledge sources.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (61876053, 62006062, 62176076, 62006060), UK Engineering and Physical Sciences Research Council (grant no. EP/V048597/1, EP/T017112/1), Natural Science Foundation of Guangdong Province of China (No. 2019A1515011705), Shenzhen Foundational Research Funding (JCYJ20200109113441941, JCYJ20210324115614039), Shenzhen Science and Technology Innovation Program (Grant No. KQTD20190929172835662), Joint Lab of Lab of HITSZ and China Merchants Securities. Yulan He is supported by a Turing AI Fellowship funded by the UK Research and Innovation (UKRI) (grant no. EP/V020579/1).

References

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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society.

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*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Erik Cambria, Yang Li, Frank Z. Xing, Soujanya Poria, and Kenneth Kwok. 2020. Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 105–114. ACM.

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*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shelly Dews and Ellen Winner. 1995. Muting the meaning a social function of irony. *Metaphor and Symbol*, 10(1):3–19.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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

Raymond W Gibbs. 2007. On the psycholinguistics of sarcasm. *Irony in language and thougt: A cognitive science reader*, pages 173–200.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

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)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Amit Kumar Jena, Aman Sinha, and Rohit Agarwal. 2020. C-net: Contextual network for sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 61–66, Online. Association for Computational Linguistics.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021a. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4707–4715. Association for Computing Machinery.

Bin Liang, Hang Su, Lin Gui, Erik Cambria, and Ruifeng Xu. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235:107643.

Bin Liang, Hang Su, Rongdi Yin, Lin Gui, Min Yang, Qin Zhao, Xiaoqi Yu, and Ruifeng Xu. 2021b. Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 208–218, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. Affective dependency graph for sarcasm detection. In *the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, pages 1844–1849.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

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*, pages 1383–1392, Online. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Information Retrieval*, 2(1-2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1136–1145.

Qiaoyu Tan, Ninghao Liu, Xing Zhao, Hongxia Yang, Jingren Zhou, and Xia Hu. 2020. Learning to hash with graph neural networks for recommender systems. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1988–1998. ACM / IW3C2.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. 2019. Learning actor relation graphs for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9964–9974. Computer Vision Foundation / IEEE.

Guo-Sen Xie, Jie Liu, Huan Xiong, and Ling Shao. 2021. Scale-aware graph neural network for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5475–5484.

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, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2115–2124. ACM.

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*, pages 3777–3786, Online. Association for Computational Linguistics.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. 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)*, pages 328–339, Online. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 974–983. ACM.

Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. 2021. Multi-modal relational graph for cross-modal video moment retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2215–2224.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578, Hong Kong, China. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14347–14355.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.