Multi-Modal Sarcasm Detection Based on Dual Generative Processes

Huiying Ma[1] , **Dongxiao He**[1] , **Xiaobao Wang**[1*] , **Di Jin**[1] , **Meng Ge**[3] and **Longbiao Wang**[1,2]
[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China
[3]Saw Swee Hock School of Public Health, National University of Singapore, Singapore
{mahuiying, hedongxiao, wangxiaobao, jindi, longbiao_wang}@tju.edu.cn, gemeng@nus.edu.sg

Abstract

With the advancement of the internet, sarcastic sentiment expression on social media has grown increasingly diverse. Consequently, multimodal sarcasm detection has emerged as a valuable tool for users to comprehend and interpret sarcastic expressions. Previous research suggests that effectively integrating three modalities (namely image, text, and their inconsistencies) enhances sarcasm detection. However, in some instances, sarcasm detection can be achieved using a single modality, while others necessitate multiple modalities for accurate recognition. This variability suggests that each modality contributes differently to sarcasm detection, and employing a traditional fusion method may introduce bias in the information, unable to explicitly demonstrate the prediction ability of each modality. Therefore, we propose a multimodal sarcasm detection method based on dual generative processes. The dual generative processes map features into the same semantic space to deeply explore emotional inconsistencies between modalities. Concurrently, by incorporating the concept of strong and weak modalities, we explicitly model the modalities' contributions based on prediction performance and autonomously adjust the weight distribution. Experimental results on publicly available multi-modal sarcasm detection datasets validate the superiority of our proposed model.

1 Introduction

Sarcasm is a prevalent form of emotional expression in daily life, often conveying intentions or viewpoints that are contrary to their literal meaning using humor or contemptuous emotions [Dews and Winner, 1995; Gibbs and Colston, 2007; Wang *et al.*, 2018]. With the flourishing development of social media, users increasingly employ rhetorical devices to express opinions or emotions online. Consequently, internet texts contain more instances of jokes, sarcasm, and humor. Such usage poses significant challenges for natural language processing, as traditional sentiment analysis struggles to accurately identify the true sentiments in sarcastic texts. Identifying sarcasm is crucial to understanding people's genuine emotions and thoughts. In the early stages of sarcasm detection, the focus was primarily on detecting syntactic patterns or special symbol tags as inherent features [Felbo *et al.*, 2017]. Other approaches involved modeling the inconsistencies within the textual modality itself [Wang *et al.*, 2023; Yu *et al.*, 2023]. With the increasing richness of online information, many user-generated contents consist of a mix of textual and visual information. Consequently, there is a growing interest in multi-modal sarcasm detection approaches [Liang *et al.*, 2022; Hua *et al.*, 2023; Dong *et al.*, 2023] that consider both textual and visual information, as they provide a more comprehensive understanding of the context.

(a) (b) (c)

Figure 1: Three scenarios of multi-modal sarcasm formation. The textual content of these posts is as follows: (a) Bye mailbox. I'm excited to go buy a new one this wknd cuz i love spending money on things other ppl break. (b) This is a good one. (c) Learn parking from this lawyer.

Many studies on multi-modal sarcasm detection have focused on learning relationships within modalities and between modalities. Some research attempts to detect sarcasm by simply fusing data from two modalities, such as using simple concatenation [Schifanella *et al.*, 2016]. Others employ attention mechanisms and external knowledge to implicitly fuse features from visual and textual modalities [Xu *et al.*, 2020; Pan *et al.*, 2020; Dong *et al.*, 2024; Zhu *et al.*, 2024]. Additionally, some studies have achieved promising results by constructing cross-modal graphs to capture features between different modalities [Liang *et al.*, 2021; Liang *et al.*, 2022; Zheng *et al.*, 2023]. Although these mod-

*Corresponding author

els have achieved remarkable performance, they still exhibit certain limitations. Often, there is ample information to identify sarcasm from just one modality, yet existing research typically assumes that both modalities' features are necessary for sarcasm detection, potentially introducing bias in the information fusion process. As illustrated in Figure 1, we can broadly delineate three scenarios: mainly relying on the emotional inconsistency information within the textual/image modality to identify sarcasm, or primarily judging through emotional inconsistency between the two modalities. In Figure 1(a), the image of a broken mailbox alone does not convey any sarcasm, but the emotional inconsistency between the phrases "very happy" and "someone else broke" serves as the primary indicator of sarcasm. Conversely, in Figure 1(b), when the text provides no useful information, the emotional inconsistency portrayed by the image becomes the primary basis for detecting sarcasm: The little girl watches table tennis on television but ignores the table tennis next to her. In scenarios where neither modality individually indicates sarcastic expression, a significant emotional inconsistency between modalities becomes the key determinant. For instance, in Figure 1(c), the strongest evidence for detecting sarcasm arises from the emotional contradiction between the phrase "learn from lawyers to park" and the contrasting image depicting a parking failure.

Drawing from the preceding insights, the process of sarcasm detection urgently requires a thorough characterization of the intricate and comprehensive effects stemming from the image modality, text modality, and their inconsistent counterparts. However, this task presents formidable challenges. Achieving it hinges on establishing minimal coupling between the learned representations of these three modalities, which entails delving deeply into the inconsistent patterns between images and text. Given the disparate semantic representation spaces of image and text modalities, direct comparison of the potential features extracted from these two modalities may inadvertently overlook their inherent differences, leading to challenges in accurately isolating the unique features of each of the three modalities. It is noteworthy that while attention mechanisms can be employed to learn the contribution of each modality, they fall short in explicitly illustrating how each modality operates and lack a certain degree of interpretability.

In this work, we propose a multi-modal sarcasm detection model based on the dual generative processes. For the exploration of the modality of inconsistency between image and text, our model learns to generate a new text feature from the image information. This generated text feature is directly compared with the encoded feature extracted from the original text modality. And similarly compared the image generated feature with the image encoded feature. By employing the dual generative processes, we map the features of different modalities to the same semantic space, facilitating joint modeling to fully learn the emotional inconsistency between image and text. Simultaneously, considering the different contribution of the three modalities to sarcasm detection, we combine the strong-weak modality mechanism [Liu *et al.*, 2018] to assess the effective information contribution of each modality based on their prediction performance, enabling the model to autonomously select the most appropriate weight distribution. The main contributions of our work are summarized as follows:

- We explicitly model the contribution of each modality to the sarcasm detection task based on their prediction performance, making the model more interpretable.
- For the deep exploration of the modality of inconsistency between image and text, we employ the dual generative processes to map the features of different modalities to the same semantic space, enabling comprehensive learning of emotional information between image and text.
- Experiments on widely used benchmark dataset show that our method outperforms state-of-the-art baselines.

2 Related Work

2.1 Multi-Modal Sarcasm Detection

The earliest sarcasm detection approaches focused on learning the inherent features of sarcastic sentences by detecting syntactic patterns or special symbols [Tay *et al.*, 2018; Lou *et al.*, 2021]. With the prevalence of multi-modal posts on social media, there has been increasing attention towards multi-modal sarcasm detection [Castro *et al.*, 2019]. [Schifanella *et al.*, 2016] first defined and addressed the task of multi-modal sarcasm detection by manually designing features. [Cai *et al.*, 2019] created a new multi-modal dataset and proposed a hierarchical model. [Xu *et al.*, 2020] modeled cross-modal comparisons and semantic correlations. [Pan *et al.*, 2020] introduced a BERT-based model that models the intra-modal and inter-modal inconsistencies. [Liang *et al.*, 2021] recognized that sarcasm information is contained in specific regions of images and certain stages of text, and proposed a graph-based approach. [Liang *et al.*, 2022] explored the use of visual objects instead of entire image regions to improve visual feature extraction. [Liu *et al.*, 2022] analyzed the information mismatches across modalities.

2.2 Generative Learning

The methodology of generative learning has found extensive applications across various domains, including image synthesis [Zhang *et al.*, 2020b], recommendation systems [Deldjoo *et al.*, 2021], and speech generation [Zhang *et al.*, 2020a]. In the context of multimodal tasks, several studies have demonstrated the efficacy of generative learning methods. [Ma *et al.*, 2019; Zellers *et al.*, 2019; Yanagi *et al.*, 2020] employed generative learning techniques to enhance multi-modal fake news detection. [Patashnik *et al.*, 2021] combined CLIP and StyleGAN to generate guided images by modifying textual input. [Frans *et al.*, 2022] utilized a pre-trained language-image encoder as a metric to maximize the similarity between descriptions and generated sketches.

3 Methodology

In this section, we provide a detailed explanation of our model, and Figure 2 illustrates the specific structure of our proposed multi-modal sarcasm detection method based on

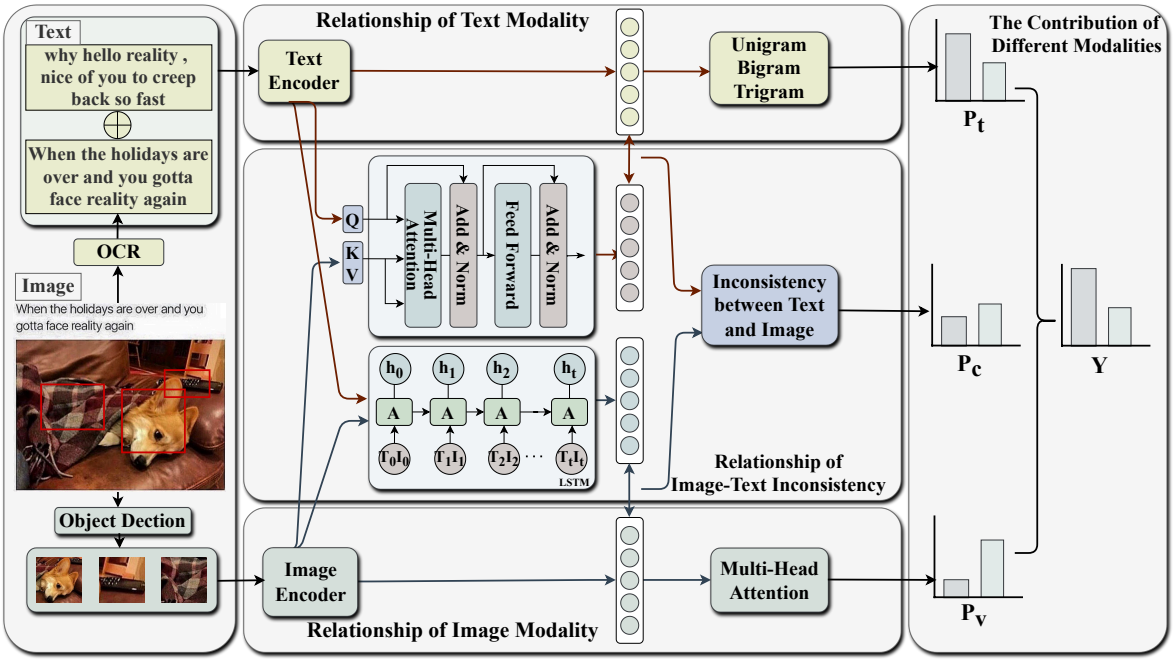

Figure 2: The overview of our proposed model.

dual generative processes (i.e., DGP). We primarily employ three components to explain our approach. Firstly, we perform multi-modal feature encoding to learn the relationships within each modality. Then, we utilize text-guided image generated feature and image-guided text generated feature to capture the emotional inconsistency between modalities. Finally, leveraging the mechanism of strong and weak modalities, we apply a multiplicative multi-modal method to assign weights to different modalities, resulting in the final detection outcome.

3.1 Problem Formulation

Giving N multi-modal samples, the i-th sample consists of three elements, i.e., $\left(X_i^T, X_i^V, Y_i\right)$, with X_i^T representing the textual content within the post, X_i^V denoting the accompanying image, and Y_i representing the true label. If the i-th sample conveys sarcastic meaning, $Y_i = 1$; otherwise, $Y_i = 0$. Our objective is to devise a model for sarcasm detection by leveraging features from both the visual and textual modalities, ultimately generating the predicted label $\hat{Y}_i$. The process is outlined as follows:

$$\mathcal{F}\left(X_i^T, X_i^V \mid \Theta\right) \longrightarrow \hat{Y}_i, \tag{1}$$

where Θ represents all the parameters of $\mathcal{F}$, and $\hat{Y}_i$ denotes the predicted results of model $\mathcal{F}$.

3.2 Multi-Modal Feature Representation

Text Embeddings

It should be noted that, apart from the given textual input X_i^T in the samples, images often contain a significant amount of embedded text in many cases. This embedded text in images often conveys important information that can capture crucial aspects of sarcasm. OCR text extraction [Pan *et al.*, 2020] has also proven successful in improving several text-related tasks. Hence, we concatenate the OCR text with the original text. The merged text is input into the pre-trained uncased BERT-base model [Devlin *et al.*, 2018] to obtain text features as follows:

$$\mathbf{T}_i = \left\{\mathbf{t}_i^1, \mathbf{t}_i^2, \ldots, \mathbf{t}_i^m\right\} = \mathrm{BERT}\left(X_i^T \oplus O_i\right), \tag{2}$$

where O_i denotes the extracted OCR text and m represents the number of tokens in the i-th merged text. After obtaining the representations of text, they are fed into a one-dimensional convolution neural network (CNN) [Singhal *et al.*, 2022] to capture the hidden local context information of sequential features, and use three window sizes (1, 2, 3, respectively) to encapsulate phrase-level information at the uni-gram, bi-gram, and tri-gram levels:

$$\mathbf{Z}_i^T = \mathrm{CNN}\left(\mathbf{W}_t \cdot \mathbf{T}_i + \mathbf{b}_t\right), \tag{3}$$

where $\mathbf{W}_t$ and $\mathbf{b}_t$ are learnable parameters in the network, $\mathbf{Z}_i^T$ represents the feature representation of the learned intra-modal relationships within the text modality.

Image Embeddings

One commonly used and classic approach to extract information from the visual modality is to perform spatially average segmentation of the complete image and then encode it. This method is widely employed in many studies as it allows for the retention of certain spatial information. However, it also preserves fragments of redundant information, either from the background or in a redundant form. Images typically contain numerous visual objects that often encompass features

related to the events or themes of the posts. In the absence of knowledge about the creator's intent, it is possible to extract more representative clues from the image, thereby distinguishing key visual objects from irrelevant ones and reducing pixel-level noise. Accurately extracting representative information from visual content can further enhance the learning of visual features. Therefore, we employ object detection toolkit [Anderson *et al.*, 2018] to extract visual objects from the images. These visual objects are then sorted based on their importance, resulting in *n* visual objects for the image $X_i^V = \{b_1, b_2, \ldots, b_n\}$. Subsequently, these visual objects are fed into the image encoder to obtain representations as:

$$\mathbf{V}_i = \left\{\mathbf{v}_i^1, \mathbf{v}_i^2, \ldots, \mathbf{v}_i^n\right\} = \mathrm{ViT}\left(X_i^V\right). \tag{4}$$

To extract individual features within the image modality, we utilize a self-attention mechanism to encode the interactions between the visual objects. This is achieved through a Transformer module that includes a multi-head self-attention layer and a position-wise feed-forward layer. Within the multi-head attention sub-layer, the attention mechanism runs in parallel multiple times. Each attention head uniquely attends to a specific portion of the sequence. Finally, all independent results are combined and linearly reshaped to obtain the desired projection size. This enables each visual object in every sample to attend to all other visual objects. The output after the multi-head self-attention module is as follows:

$$\begin{cases} Multihead = [\mathbf{head}_1 \otimes, \ldots, \mathbf{head}_j]\, \mathbf{W}_v, \\ \mathbf{head}_j = \mathrm{attention}\left(\mathbf{V}_i\mathbf{W}_j^Q, \mathbf{V}_i\mathbf{W}_j^K, \mathbf{V}_i\mathbf{W}_j^V\right)\mathbf{W}_v, \\ \mathrm{attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \mathrm{softmax}\left(\frac{\mathbf{K}\mathbf{Q}^T}{\sqrt{d_k}}\right)\mathbf{V}, \end{cases} \tag{5}$$

where **Q**, **K**, and **V** represent queries, key-value pairs, and all the W denote learnable parameter matrices. The position-wise feed-forward sub-module is applied independently and identically to each image object. Then, position-wise feed-forward and layer normalization are applied to obtain a set of continuous representations. Residual connections are applied separately in both sub-layers, resulting in the outputs $\mathbf{V}'$ and $\mathbf{Z}^V$ as follows:

$$\begin{cases} \mathbf{V}_i' = \mathrm{LN}\left(\mathbf{V}_i + \mathrm{Multihead}\left(\mathbf{V}_i\right)\right), \\ \mathbf{Z}_i^V = \mathrm{LN}\left(\mathbf{V}_i' + \mathrm{Feedforward}\left(\mathbf{V}_i'\right)\right), \end{cases} \tag{6}$$

where $\mathbf{V}'$ represents the output of image features after undergoing multi-head self-attention modules and layer normalization, while $\mathbf{Z}^V$ signifies its output after passing through position-wise feed-forward sub-layers and layer normalization.

3.3 The Inconsistency Between Text and Image With Generated Feature

In social media, there is often a discrepancy between the visual relationships expressed in images and the textual content. Inspired by [Shang *et al.*, 2022], to better capture the emotional inconsistencies between these two modalities and overcome the inherent differences between visual and textual modalities, a direct approach is to employ image synthesis tools to generate new images based on the textual modality. However, text on social media platforms typically consists of a limited number of words, and the conveyed emotions can be diverse. Thus, it may not be feasible to generate images that fully align with the semantic and contextual information of the given text. Additionally, the quality of images on social media varies, often containing pixel-level noise. Therefore, we propose a dual-generation approach, where each modality learns features from the other modality and generates features that incorporate the characteristics of the opposite modality. This approach allows us to explore the emotional inconsistencies between text and images by excluding potential influences such as dimensions in a more comprehensive manner.

We first attempt to learn image generation features guided by textual features using a Long Short-Term Memory (LSTM) network guided by image objects. The textual features are replicated to match the number of image objects and paired with the corresponding image object features, represented as $\widetilde{\mathbf{T}_i\mathbf{V}_i} = \left\{\widetilde{\mathbf{t}_i^1\mathbf{v}_i^1}, \widetilde{\mathbf{t}_i^2\mathbf{v}_i^2}, \ldots, \widetilde{\mathbf{t}_i^n\mathbf{v}_i^n}\right\}$. Then the generated image feature $\mathbf{V}_{gen}$ obtained by feeding the concatenated features into the network is as follows:

$$\mathbf{V}_i^{gen} = \mathrm{LSTM}\left(\widetilde{\mathbf{T}_i\mathbf{V}_i}\right). \tag{7}$$

By doing so, we can learn the relationships between each image-text pair. These image features capture relevant semantic relationships and other characteristics related to the textual features, enabling the generation of image features that incorporate textual characteristics. If there is a significant difference between the images and the text, it is reasonable to infer that the generated image features guided by textual features will differ significantly from the original image encoding features. Thus, the similarity between the generated image features and the encoding features can be used to represent the emotional inconsistency between the visual and textual modalities.

To generate textual features guided by images, we employ a self-attention encoder based on image objects to learn the internal relationships between each pair of image objects, represented as

$$\mathbf{V}_i = \mathrm{SelfAttention}\left(\mathbf{V}_i\mathbf{W}^Q, \mathbf{V}_i\mathbf{W}^K, \mathbf{V}_i\mathbf{W}^V\right). \tag{8}$$

After encoding the pairwise relationships between image objects, we utilize a decoder based on cross-modal multi-head attention to decode text features guided by visual features, as follows:

$$\mathbf{T}_i^{gen} = \mathrm{Multihead}\left(\mathbf{T}_i\mathbf{W}^Q, \mathbf{V}_i\mathbf{W}^K, \mathbf{V}_i\mathbf{W}^V\right), \tag{9}$$

where $\mathbf{T}_i^{gen}$ represents the text generated feature with incorporated image information. If there is a significant difference between the images and the text, it is reasonable to infer that the similarity between the generated textual features and the encoding features can be used to represent the emotional inconsistency between the visual and textual modalities. Therefore, the features of the modality of inconsistency between images and text can be represented as:

$$\mathbf{Z}_i^C = \mathrm{Sim}\left(\mathbf{V}_i^{enc}, \mathbf{V}_i^{gen}\right) + \mathrm{Sim}\left(\mathbf{T}_i^{enc}, \mathbf{T}_i^{gen}\right), \tag{10}$$

where $\mathbf{V}_i^{enc}$ and $\mathbf{T}_i^{enc}$ represent the encoded features $\mathbf{V}_i$ and $\mathbf{T}_i$ obtained through the respective encoders.

3.4 The Contribution of Different Modalities

In this task, based on the three cases mentioned earlier, we can conclude that all three modalities can contribute to sarcasm detection. In addition to the commonly used textual and visual modalities, the modality that represents the inconsistency between text and image emotions can serve as a third modality to aid in better sarcasm detection.

To utilize the information from both modalities, a straightforward approach is to concatenate the features from different modalities, which has been employed in many studies [Atrey *et al.*, 2010; Ngiam *et al.*, 2011; Neverova *et al.*, 2015; Wöllmer *et al.*, 2010]. However, this fusion method has certain limitations as it assumes that all modalities contribute equally, making it difficult to determine which modality contributes more in the sarcasm detection task. Inspired by [Singhal *et al.*, 2022], for an ideal model given multiple input modalities, it should be robust to the noise from less contributing modalities and able to extract more informative features from stronger modalities in each sample to better aid sarcasm detection. Based on the assumption that the contribution from each modality is not equal, we employ a multiplication-based multi-modal method [Liu *et al.*, 2018]. Each modality independently makes predictions based on its own feature information in each sample as follows:

$$\begin{cases} \mathbf{P}_t = \mathrm{MLP_t}\left(\mathbf{Z}_i^T\right) = \left[p_t^1, p_t^0\right], \\ \mathbf{P}_v = \mathrm{MLP_v}\left(\mathbf{Z}_i^V\right) = \left[p_v^1, p_v^0\right], \\ \mathbf{P}_c = \mathrm{MLP_c}\left(\mathbf{Z}_i^C\right) = \left[p_r^1, p_r^0\right], \end{cases} \tag{11}$$

where $\mathbf{P}_t$, $\mathbf{P}_v$ and $\mathbf{P}_c$ represent the predictions for text, image, and the inconsistency between image and text, respectively. The typical additive combination involves summing the individual loss functions of all modalities:

$$L^y_{crossentropy} = -\sum_{k=1}^{X} \log\left(p_k^y\right), \tag{12}$$

where L^y represents the loss of the category, X represents the number of modalities involved in the task, and k represents the index of these modalities. To better capture the contributions of each modality in sarcasm detection, we incorporated a weighting factor H_k for the k-th modality into our approach:

$$H_k = \left[\prod_{j \neq k}\left(1 - p_j^y\right)\right]^{\delta/(X-1)}, \tag{13}$$

where the hyper-parameter δ is used to control the extent of learning for the weighting factor, which determines the degree of utilization of the multi-modal data. If a particular modality demonstrates better predictive performance, the relative weight H_k of that modality will be larger, thereby highlighting the influence of the strong modality. The loss function for the multiplication-based multi-modal method can be defined as follows:

$$L^y_{multiplicative} = -\sum_{k=1}^{X} H_k \cdot \log\left(p_k^y\right). \tag{14}$$

	Training	Development	Testing
Sarcasm	8642	959	959
Non-Sarcasm	11174	1451	1450
All	19816	2410	2409

Table 1: Statistics of the experimental data.

4 Experiments

In this section, we begin by presenting the necessary preparations for the experiments, including the dataset, experimental setup, and comparative models. After analyzing the experimental results, we proceed with conducting ablation experiments and case studies to further enhance our understanding of the effectiveness of the proposed model.

4.1 Dataset

We evaluate our model on the publicly available multi-modal sarcasm detection benchmark dataset collected from Twitter by [Cai *et al.*, 2019]. Each sample in the dataset consists of an English tweet accompanied by an associated image. Samples identified as expressing sarcasm are considered positive examples, while non-sarcastic expressions are considered negative examples. The dataset is divided into training, validation, and test sets, with an approximate ratio of 80%:10%:10%. Table 1 provides specific statistics regarding the dataset.

4.2 Experimental Settings

To ensure a fair comparison, we follow the preprocessing steps described in [Xu *et al.*, 2020; Liang *et al.*, 2021]. We remove samples containing frequently co-occurring words with sarcastic expressions (e.g., "irony", "sarcasm") after preprocessing, to avoid introducing external information. We utilize the Huggingface pre-trained BERT-base-uncased model [Devlin *et al.*, 2018] for encoding, representing each text as a 768-dimensional vector. Regarding images, when the number of extractable objects on each image is indeterminate, we select the top 20 objects with the highest scores. These objects are resized to a uniform size of 224×224, and we employ the pre-trained ViT model [Dosovitskiy *et al.*, 2020] to encode each object into a 768-dimensional embedding. We employ the Adam optimizer with a learning rate of 1e-4. To prevent over-fitting, we apply a dropout rate of 0.1 for intra-modal learning, and early stopping techniques are utilized. The parameters and batch size are set to 0.8 and 32, respectively.

Following [Cai *et al.*, 2019], we evaluate the model's performance using Accuracy, Precision, Recall, and F1-score. Due to the imbalanced label distribution in the dataset, as suggested in [Pan *et al.*, 2020], we also employ macro metrics to assess the model. The experimental results of our proposed model are obtained by running 10 iterations with random initialization and then averaging the obtained results.

4.3 Baselines

We compare our proposed model with several state-of-the-art baseline models, which can be broadly categorized into:

(1) Models Based on Image Modality: These models solely utilize image information for sarcasm detection. They include **Image** [Cai *et al.*, 2019], which classifies images

based on the vectors obtained after pooling layers in ResNet [He *et al.*, 2016]; **ViT** [Dosovitskiy *et al.*, 2020] which divides images into multiple patches and utilizes the [CLS] token for sarcasm detection.

(2) Models Based on Text Modality: These models solely utilize text information for sarcasm detection. They include the following models: **TextCNN** [Kim, 2014]: A deep learning model based on convolutional neural networks (CNN) that captures n-gram features for text classification tasks; **Bi-LSTM** [Graves and Schmidhuber, 2005]: Bi-directional LSTM network that leverages bidirectional processing to learn text representations and performs predictions using a classification layer; **SIARN** [Tay *et al.*, 2018]: A model that incorporates an internal attention mechanism for text sarcasm detection; **SMSD** [Xiong *et al.*, 2019]: It captures inconsistent information in the text by considering the interactions between words in a sentence; **BERT** [Devlin *et al.*, 2018]: It directly encodes text by adding a "[CLS] text [SEP]" token and generates predictions based on the representations.

(3) Models Based on Multi-Modality: These models leverage both text and image modalities for sarcasm detection. They include: **HFM** [Cai *et al.*, 2019]: A hierarchical model that performs multi-modal feature fusion by extracting attribute words of images; **D&RNet** [Xu *et al.*, 2020]: It models text-context comparisons and cross-modal semantic correlations for sarcasm detection; **Att-BERT** [Pan *et al.*, 2020]: It explores the inconsistency in multi-modal sarcasm detection using cross-modal attention and a co-attention of intra-modality; **InCrossMGs** [Liang *et al.*, 2021]: It detects sarcasm by exploring sarcasm information within and across modalities using graph convolutional networks; **CMGCN** [Liang *et al.*, 2022]: It captures inconsistencies between modalities based on graph convolutional networks using object detection; **HCM** [Liu *et al.*, 2022]: It performs sarcasm detection by leveraging atomic-level and compositional-level consistency between images and text; **ERGCN** [Li *et al.*, 2023]: It enables the extraction of valuable entity information from visual objects in images and textual descriptions, and leverages external knowledge to construct cross-modal graphs for each image-text pair sample to facilitate the identification of semantic inconsistencies between modalities.

4.4 Main Experimental Results

The compared results are summarized in Table 2, and we draw the following conclusions: (1) Our proposed model outperforms the state-of-the-art models in terms of all metrics, achieving 0.96% improvement in terms of accuracy and 4.78% improvement in terms of F1 score. This demonstrates that our model, utilizing generated features and the concept of strong and weak modalities, significantly enhances the performance of multi-modal sarcasm detection. (2) Compared to the Att-BERT, which utilizes attention mechanisms to explore inconsistencies within and between modalities, our model achieves higher performance. This confirms the effectiveness of utilizing generated features to explore the inter-modal emotional inconsistencies. (3) Compared to InCrossMGs, CMGCN, and HCM, which fuse multi-modal features, our model demonstrates superiority, highlighting the improvement in sarcasm detection through the utilization of a multiplicative multi-modal approach. (4) Furthermore, the text-based approach consistently outperforms the image-based approach across all metrics, indicating that there may be more information in the text modality that aids in sarcasm detection. This also confirms our hypothesis that "not all modalities contribute equally to sarcasm detection". (5) However, while the text-based model significantly outperforms the image-based model, it still falls behind the models that consider both image and text modalities. This suggests that incorporating information from both image and text modalities remains more effective for sarcasm detection overall.

Model		Acc(%)	P(%)	R(%)	F1(%)
Image	Image	64.76	54.41	70.80	61.53
	ViT	67.83	57.93	70.07	63.43
Text	TextCNN	80.03	74.29	76.39	75.32
	Bi-LSTM	81.90	76.66	78.42	77.53
	SIARN	80.57	75.55	75.70	75.63
	SMSD	80.90	76.46	75.18	75.82
	BERT	83.85	78.72	82.27	80.22
Both	HFM	83.44	76.57	84.15	80.18
	D&R Net	84.02	77.97	83.42	80.60
	Att-BERT	86.05	77.80	84.15	80.85
	InCrossMGs	86.10	81.38	84.36	82.84
	CMCGN	84.94	79.68	83.44	81.52
	HCM	86.38	86.75	79.50	82.79
	ERGCN	86.72	82.57	84.46	83.51
	DGP	**87.21**	**87.10**	**86.48**	**86.75**

Table 2: Comparison of experimental results on the publicly available dataset. The best results are in bold.

Model	**Acc(%)**	**F1(%)**
w/o $multiplicative$	85.09	84.62
w/o $generated$	85.21	84.57
w/o $text_{gen}$	85.01	84.42
w/o $image_{gen}$	85.38	84.98
w/o $text$	84.63	84.21
w/o $image$	84.93	84.34
DGP	**87.21**	**86.75**

Table 3: Experimental results of ablation study.

4.5 Ablation Study

To demonstrate the effectiveness of each module in our proposed method, the results are presented in Table 3, we conduct several ablation experiments comparing our model with different variants: 1) **w/o** $multiplicative$: This variant removes the multiplicative multi-modal method we employed and replaces it with a simple summation of predictions from different modalities, demonstrating the significant performance improvement achieved by the strong and weak modality mechanism for multi-modal sarcasm detection. 2) **w/o** $generated$: This variant removes the cross-modal correlation learning using generated features, focusing solely on the intra-modal learning, confirming the importance of learning emotional inconsistencies across modalities for sarcasm detection. 3) **w/o** $text_{gen}$ and **w/o** $image_{gen}$: These vari-

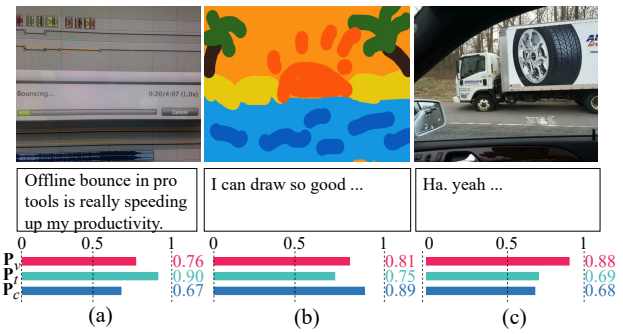

Figure 3: Three cases of sarcasm detected by our proposed model. $\mathbf{P}_v$, $\mathbf{P}_t$, and $\mathbf{P}_c$ represent the relative prediction values assigned to the image, text and the image-text inconsistency, respectively.

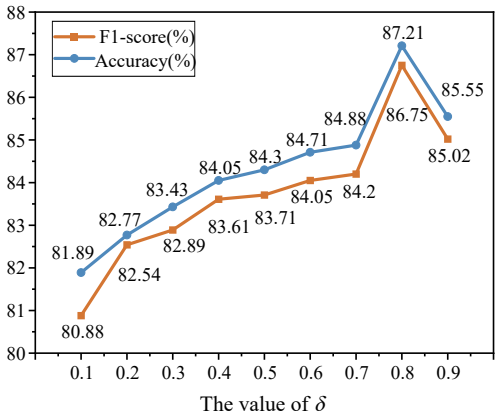

Figure 4: The impact of hyper-parameter δ magnitudes on the performance of our proposed model.

ants remove the text-guided and image-guided feature generation, respectively, which not only demonstrates their contributions to the task but also highlights that the removal of text-generated features results in a more significant performance drop compared to removing image-generated features. This implies that the text-generated features can capture more information. 4) **w/o** *text* and **w/o** *image*: These variants remove the intra-modal learning for text or image respectively, leading to varying degrees of performance degradation, further confirming that each modality contributes differently to sarcasm detection and validating the effectiveness of the strong and weak modality mechanism.

4.6 Case Study

To provide a more intuitive understanding of how our proposed model identifies the essential components of sarcasm, we conduct a qualitative analysis. Figure 3 displays the prediction values $\mathbf{P}_v$, $\mathbf{P}_t$, and $\mathbf{P}_c$ of each modality across various scenarios, elucidating their individual contributions to the task of sarcasm detection.

In Figure 3(a), the text juxtaposes the portrayal of difficulty with an emphasis on expediting work efficiency, forming an emotionally sarcastic context. The prediction values assigned to the image, text, and modal inconsistency are 0.76, 0.9, and 0.67, respectively, underscoring the text's substantial contribution. In Figure 3(b), the text showcases exceptional drawing skills, while the image depicts crude and abstract artwork, leading to an emotional inconsistency. The prediction values assigned to the three modalities are 0.81, 0.75, and 0.89, respectively. Here, the inconsistency between modalities contributes the most relevant information, although both the image and text modalities also make some contributions. In Figure 3(c), the image primarily conveys the sarcastic sentiment through the contrast between a tire advertisement and a depiction of a punctured tire. The prediction values for the three modalities are 0.88, 0.69, and 0.68, respectively, highlighting the predominant role of the image modality.

4.7 The Influence of δ on the Model

In this section, we analyze the impact of hyper-parameter δ magnitudes on the performance of our proposed model, which can represent the extent of change in the weights of each modality during updates. We set the values of δ ranging from 0.1 to 0.9 and present the results in Figure 4. It can be observed that the model architecture performs better when δ is set to 0.8 compared to other configurations. As δ decreases, the model's performance deteriorates. This suggests that the weights of each modality change to a lesser degree during the updating process, indicating insufficient adjustment of the relative importance among modalities. On the other hand, in the case where δ is 0.9, the model's performance does not reach its optimal performance. This may be attributed to a significant decrease in weights, which simultaneously reduces the importance of certain modalities and ultimately diminishes the model's overall performance.

5 Conclusion

In this paper, we propose a multi-modal sarcasm detection method based on dual generative processes. Our approach comprises two primary components. Firstly, recognizing the varying effectiveness of information contribution from different modalities, we introduce a multiplicative multi-modal approach grounded in the concept of strong and weak modalities. By dynamically adjusting the degree of variation in modality weights, we automatically adapt to the most appropriate weight distribution, thereby enhancing the performance of the multi-modal sarcasm detection task. Secondly, acknowledging that the feature representations of image and text modalities reside in distinct semantic spaces, we leverage a dual-generation process to deeply explore the information from the modality inconsistency between image and text. By utilizing the features of one modality to guide the generation of features from the other modality, we mitigate the risk of overlooking the inherent semantic differences between images and text. Extensive experiments conducted on publicly available datasets demonstrate the effectiveness and superiority of our proposed method, offering valuable insights for research in the field of multi-modal sarcasm detection.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2023YFC3304503), National Natural Science Foundation of China (62302333, 92370111, 62276187, 62272340, U23B2053) and the China Postdoctoral Science Foundation (Grants No. 2023M732593).

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[Atrey *et al.*, 2010] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379, 2010.

[Cai *et al.*, 2019] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 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, 2019.

[Castro *et al.*, 2019] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*, 2019.

[Deldjoo *et al.*, 2021] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[Dong *et al.*, 2023] Yiqi Dong, Dongxiao He, Xiaobao Wang, Yawen Li, Xiaowen Su, and Di Jin. A generalized deep markov random fields framework for fake news detection. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pages 4758–4765, 2023.

[Dong *et al.*, 2024] Yiqi Dong, Dongxiao He, Xiaobao Wang, Youzhu Jin, Meng Ge, Carl Yang, and Di Jin. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. 2024.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Felbo *et al.*, 2017] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

[Frans *et al.*, 2022] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35:5207–5218, 2022.

[Gibbs and Colston, 2007] Raymond W Gibbs and Herbert L Colston. *Irony in language and thought: A cognitive science reader*. Psychology Press, 2007.

[Graves and Schmidhuber, 2005] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Hua *et al.*, 2023] Jiaheng Hua, Xiaodong Cui, Xianghua Li, Keke Tang, and Peican Zhu. Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136:110125, 2023.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[Li *et al.*, 2023] Lingshan Li, Di Jin, Xiaobao Wang, Fengyu Guo, Longbiao Wang, and Jianwu Dang. Multi-modal sarcasm detection based on cross-modal composition of inscribed entity relations. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence*, pages 918–925. IEEE, 2023.

[Liang *et al.*, 2021] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 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, 2021.

[Liang *et al.*, 2022] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng 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)*, pages 1767–1777, 2022.

[Liu *et al.*, 2018] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.

[Liu *et al.*, 2022] Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. *arXiv preprint arXiv:2210.03501*, 2022.

[Lou *et al.*, 2021] Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. Affective dependency graph for sarcasm detection. In *Proceedings of the 44th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1844–1849, 2021.

[Ma *et al.*, 2019] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*, pages 3049–3055, 2019.

[Neverova *et al.*, 2015] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 689–696, 2011.

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

[Patashnik *et al.*, 2021] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

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

[Shang *et al.*, 2022] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A duo-generative approach to explainable multimodal covid-19 misinformation detection. In *Proceedings of the ACM Web Conference*, pages 3623–3631, 2022.

[Singhal *et al.*, 2022] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference*, pages 726–734, 2022.

[Tay *et al.*, 2018] Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. Reasoning with sarcasm by reading in-between. *arXiv preprint arXiv:1805.02856*, 2018.

[Wang *et al.*, 2018] Zhen Wang, Marko Jusup, Lei Shi, Joung-Hun Lee, Yoh Iwasa, and Stefano Boccaletti. Exploiting a cognitive bias promotes cooperation in social dilemma experiments. *Nature Communications*, 9(1):2954, 2018.

[Wang *et al.*, 2023] Xiaobao Wang, Yiqi Dong, Di Jin, Yawen Li, Longbiao Wang, and Jianwu Dang. Augmenting affective dependency graph via iterative incongruity graph learning for sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4702–4710, 2023.

[Wöllmer *et al.*, 2010] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth S. Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Interspeech*, pages 2362–2365, 2010.

[Xiong *et al.*, 2019] Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *Proceedings of the ACM Web Conference*, pages 2115–2124, 2019.

[Xu *et al.*, 2020] Nan Xu, Zhixiong Zeng, and Wenji 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*, pages 3777–3786, 2020.

[Yanagi *et al.*, 2020] Yuta Yanagi, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga. Fake news detection with generated comments for news articles. In *2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, pages 85–90. IEEE, 2020.

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

[Zellers *et al.*, 2019] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.

[Zhang *et al.*, 2020a] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

[Zhang *et al.*, 2020b] Yang Zhang, Xiangyu Dong, Md Tahmid Rashid, Lanyu Shang, Jun Han, Daniel Zhang, and Dong Wang. Pqa-cnn: Towards perceptual quality assured single-image super-resolution in remote sensing. In *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*, pages 1–10. IEEE, 2020.

[Zheng *et al.*, 2023] Yizhen Zheng, He Zhang, Vincent Lee, Yu Zheng, Xiao Wang, and Shirui Pan. Finding the missing-half: Graph complementary learning for homophily-prone and heterophily-prone graphs. In *International Conference on Machine Learning*, pages 42492–42505. PMLR, 2023.

[Zhu *et al.*, 2024] Peican Zhu, Botao Wang, Keke Tang, Haifeng Zhang, Xiaodong Cui, and Zhen Wang. A knowledge-guided graph attention network for emotion-cause pair extraction. *Knowledge-Based Systems*, 286:111342, 2024.