

Towards multimodal sarcasm detection via label-aware graph contrastive learning with back-translation augmentation[☆]

Yiwei Wei^a, Maomao Duan^b, Hengyang Zhou^b, Zhiyang Jia^b, Zengwei Gao^b, Longbiao Wang^{a,*}

^a Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China

^b College of Petroleum Engineering, China University of Petroleum(Beijing) at Karamay, Karamay, China

ARTICLE INFO

Keywords:

Multimodal sarcasm detection
Label-aware contrastive learning
Back-translation augmentation

ABSTRACT

Multimodal sarcasm detection, as a sentiment analysis task, has witnessed great strides owing to the rapid development of multimodal machine learning. However, existing graph-based studies mainly focus on capturing the atomic-aware relations between textual and visual graphs within individual instances, neglecting label-aware connections between different instances. To address this limitation, we propose a novel Label-aware Graph Contrastive Learning (LGCL) method that detects ironic cues from a label-aware perspective of multimodal data. We first construct unimodal graphs for each instance and fuse them into graph semantic space, to obtain the multimodal graphs. Then, we introduce two label-aware graph contrastive losses: Label-aware Unimodal Contrastive Loss (LUCL) and Label-aware Multimodal Contrastive Loss (LMCL), to make the model aware of the shared ironic cues related to sentiment labels within multimodal graph representations. Additionally, we propose Back-translation Data Augmentation (BTrA) for both textual and visual data to enhance contrastive learning, where different back-translation schemes are designed to generate a larger number of positive and negative samples. Experimental results on two public datasets demonstrate our method achieves state-of-the-art (SOTA) compared to previous methods.

1. Introduction

Sarcasm, a commonly used form of figurative expression, is utilized in various real-life situations to convey meanings that are opposite to their literal interpretations [1]. It plays a crucial role in analyzing human sentiment and perspective within conversations, steadily offering benefits across a spectrum of domains. These domains include natural language dialogue [2], public sentiment detection [3], and meticulous analysis of the nuanced dynamics within social media [4]. With the widespread use of mobile internet and smartphones, increasing users are willing to post multimodal data (text, images, videos) to express their feelings and sentiments on various topics. As a result, multimodal sarcasm detection has gained increasing research attention in recent years [5–9], emerging as a highly sought-after topic in the field of natural language processing and multimedia computing.

Till now, extensive research has been conducted to tackle the task of multimodal sarcasm detection. Previous studies, such as Schifanella et al. [10] and MMSD [5], have explored the fusion of different feature vectors to combine multimodal features. Nevertheless, solely relying

on multimodal data fusion is insufficient to tackle this task, as the key to effective multimodal sarcasm detection lies in accurately extracting incongruent sentiment cues from different modalities. Consequently, some methods [9,11,12] attempt to model this characteristic of incongruity between image and text with the attention mechanisms, optimal transport method, and dynamic network. To model the relationship between modalities more accurately, the following studies [7,8,13] introduce graph-based methods to capture the incongruity between modalities data. Although promising, current graph-based methods mainly consider modeling the incongruity between visual and textual graphs within a single instance, while neglecting the common irony characteristics exhibited by the instances that share the same label. For example, Fig. 1 illustrates four image-text pairs labeled as sarcasm convey similar sarcastic cues, where visual regions contradict the meaning of phrases (e.g. lovely weather, great weather) in the text. This inspired us to utilize sentimental labels to draw intricate connections among instances, thereby facilitating learning sarcastic clues.

In this paper, we propose a novel Label-aware Graph Contrastive Learning (LGCL) method that involves several key components. First,

[☆] This work is supported by the First Karamay City Science and Technology Innovation Talent - Core Science and Technology Innovation Talent and the National Natural Science Foundation of China (No. XQZX20230110).

* Corresponding author.

E-mail addresses: duanmm1990@126.com (M. Duan), hengyangzhou@outlook.com (H. Zhou), jiazhongyang@cupk.edu.cn (Z. Jia), 2020592211@cupk.edu.cn (Z. Gao), longbiao_wang@tju.edu.cn (L. Wang).

<https://doi.org/10.1016/j.knosys.2024.112109>

Received 18 December 2023; Received in revised form 5 May 2024; Accepted 9 June 2024

Available online 27 June 2024

0950-7051/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

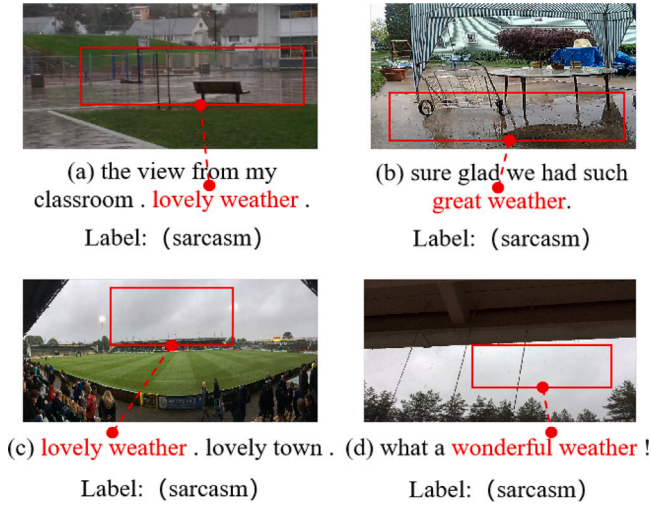


Fig. 1. Four multi-modal sarcastic examples. Boxes and words in the red color denote highly correlated sarcastic cues.

we independently encode the visual graph and textual graph with graph attention network (GAT) [14] to capture the unique characteristics of each modality. Additionally, we introduce a multimodal graph fusion module that combines the two unimodal graphs into a unified graph semantic space. Thus, we can leverage multimodal complementary strengths and enhance the overall representation. Moreover, we propose two label-aware graph contrastive losses. The first loss called the label-aware unimodal contrastive loss (LUCL), operates on the graph representations from the unimodal encoders. It treats the unimodal graphs within an instance as positive samples, encouraging them to be similar while considering the graphs from instances with different labels as negative samples. This loss promotes intra-instance multimodal alignment by creating a clear separation in the negative sample space. The second loss called label-aware multimodal contrastive loss (LMCL), is applied after the graph fusion process. It focuses on multimodal graphs within the same label, considering them as positive samples, while treating graphs from instances with different labels as negative samples. By applying this loss, our model learns to generate more similar multimodal graph representations for instance pairs with the same label, while pushing away those pairs that have different labels. To facilitate contrastive learning, we propose a data augmentation method called Back-translation Augmentation (BTrA). This method generates a larger number of positive and negative samples for both text and image modalities. For text, we leverage pre-trained large language model GPT-3.5 [15], to translate and back-translate existing text, resulting in augmented text. For image augmentation, we employ a pre-trained BLIP model [16] to generate captions for images, which are then used as prompts for the pre-trained stable-diffusion model [17] to translate back into augmented images. Remarkably, this is the first attempt to employ the large language model for data augmentation in sentiment detection tasks. For evaluation, we conducted experiments on the publicly available multimodal sarcasm detection datasets, MMSD [5] and MMSD2.0 [18]. The results demonstrate that our model achieves state-of-the-art performance across all metrics.

1. We propose a novel Label-aware Graph Contrastive Learning (LGCL) method for multi-modal sarcasm detection. It aims to overcome the limitation that existing multi-modal sarcasm detection models only consider modeling the incongruity between visual and textual graphs within individual instances, while neglecting the common ironic characteristics among instances with the same labels.

2. To improve the performance of contrastive learning, we propose Back-translation Augmentation, leveraging different back-translation methods. This approach allows us to generate a larger number of positive and negative samples. Notably, our research is the first attempt to employ the large language model for data augmentation in sentiment detection task.
3. Experimental results show our model achieves state-of-the-art (SOTA) on two public multimodal sarcasm detection datasets MMSD [5] and MMSD2.0 [18] across all metrics. We also conduct extensive experiments to validate the effectiveness and contribution of the different components in our proposed method.

2. Related work

2.1. Multimodal sarcasm detection

Multimodal sarcasm detection has emerged as a progressively challenging task due to the escalating demand for analyzing multimodal content across social media platforms. Schifanella et al. [10] pioneered this field by addressing it as a multimodal classification challenge, merging manually engineered multimodal features. Subsequently, MMSD [5] proposed a hierarchical fusion model that amalgamates features from textual and visual modalities, leveraging a novel multimodal sarcasm detection dataset rooted in Twitter data. D&R Net [19] introduced the Decomposition and Relation Network, capturing contextual contrasts and semantic associations within multimodal data. Att-BERT [20] harnessed co-attention and self-attention mechanisms for imbibing congruity information within and between modalities. DynRT-Net [21] modeled the dynamic mechanism to restrict the model from dynamically adjusting to diverse image-text pairs. Within the domain of graph-based methodologies, InCrossMGs [6] probed sentiment inconsistencies through intra- and cross-modal graph construction. CMGCN [13] conceived cross-modal graphs to delineate ironic relationships between textual and visual elements. Moreover, HKEmodel [7] introduced hierarchical congruity modeling via cross-attention mechanisms and graph neural networks. Meanwhile, MILNet [8] orchestrated three distinct graphs to decipher local and global incongruities. Although promising, existing graph-based methods mainly focus on how to draw multimodal graphs and integrate graph features, while neglecting to draw the intricate connections among instances based on sentiment labels from existing instances to facilitate prediction.

2.2. Graph neural networks

Graph Neural Networks (GNNs) have emerged as a transformative paradigm for capturing intricate relationships within graph-structured data. The pioneering work introduced Graph Convolutional Networks (GCNs) [22], propelling graph analysis by fusing spectral graph theory with neural networks. This catalyzed a wave of innovations, including Graph Attention Networks (GAT) [14], which elegantly incorporate attention mechanisms to selectively aggregate node information. Meanwhile, GraphSAGE [23] champions inductive learning, adeptly learning representations from diverse nodes by neighborhood aggregation. Recent strides have showcased the versatility of GNNs across domains. Sui et al. [24] leveraged GNNs for causal inference in graphs, unveiling underlying causal structures. Link prediction was reinvigorated by Kou et al. [25], using GNNs, illuminating evolving network dynamics. GNNs have also redefined recommendation systems [26], personalizing recommendations by intertwining graph insights. Despite their successes, the potential of GNNs for multimodal modeling tasks can be further explored.

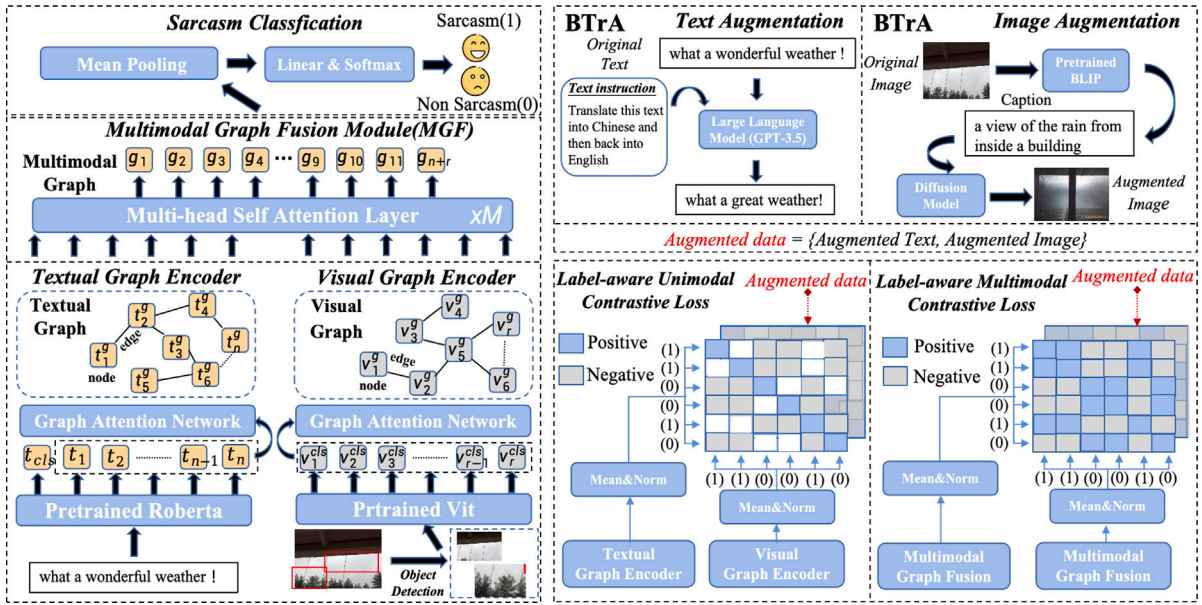


Fig. 2. The overall framework of the proposed method for multimodal sarcasm detection. In the framework, the left part denotes the Multimodal Graph Fusion Process. While the bottom right part denotes the Label-aware Graph Contrastive Learning (LGCL) and the top right part denotes the Back-translation Data Augmentation (BTrA).

2.3. Contrastive learning

Contrastive learning, aiming to learn meaningful representations for a given task, has been widely applied in various fields, such as natural language processing (NLP), computer vision, and multimodal domains. In NLP, ConSERT [27], Sim-CSE [28], and CLEAR [29] have employed contrastive learning to capture discriminative semantic information. In computer vision, SimCLR [30], Sim-Siam [31] and CLIP [32] have applied this technique to improve the performance of the model. Recently, contrastive learning has also gained popularity in recommendation systems [33–36]. In particular, GraphCL [34], GCA [35], and DSGC [36] have applied graph contrastive learning with graph augmentations for graph neural networks to handle data heterogeneity in graphs. However, graph contrastive learning has not been explored in the field of multi-modal sarcasm detection, and meanwhile, those methods relying on designing a data augmentation in the structural domain to enlarge contrastive pairs are not adapted to multi-modal sarcasm detection.

3. Methodology

In this section, we introduce the details of the LGCL method. The overall architecture of the LGCL is illustrated in Fig. 2, which mainly consists of three components: Multimodal Graph Fusion (MGF) Process, Label-aware Graph Contrastive Learning (LGCL), and Back-translation Augmentation (BTrA).

3.1. Multimodal graph fusion process

The left of Fig. 2 depicts the multimodal graph fusion process, including two unimodal encoders and a multimodal graph fusion module. The image and text inputs are first encoded by unimodal graph encoders respectively to obtain the unimodal graphs. Then, we feed the unimodal graphs into a multimodal graph fusion module that is composed of M stacked multi-head self-attention layers [37] to learn cross-modal incongruity, resulting in multimodal graphs with cross-modal information.

Textual Graph Encoder. Given a sequence of words $S = \{w_i\}_{i=1}^n$, where n represents the length of the text S , we utilize the pre-trained

model RoBERTa [38] as the text encoder. The RoBERTa model encodes each word w_i into a d^T -dimensional embedding as follows:

$$X^t = [t_{cls}, t_1, t_2, \dots, t_n] = \text{RoBERTa}([CLS, w_1, w_2, \dots, w_n]) \quad (1)$$

Here, X^t is the textual representation of the input text, and CLS denotes the class token, which is excluded from the subsequent graph construction. To unify the dimensions of representations across different modalities, we add an additional multi-layer perceptron (MLP) following the text encoder to transform the dimension d^T to d . Next, we follow the approach presented in the previous work [13] to construct the textual graph. In this approach, the input tokens are modeled as graph nodes, while the dependency relations between words, obtained through the spaCy library,¹ are utilized as the graph edges. Concretely, an edge between two words is established in the textual graph whenever a dependency relation is detected between them. After that, the textual graph is fed into 2-layer graph attention networks (GAT) to obtain the final representation, which can be defined as:

$$\alpha_{ij}^l = \frac{\exp\left(\text{LeakyReLU}\left(\bar{a}_i^T [W_l^T \bar{t}_i || W_l^T \bar{t}_j]\right)\right)}{\sum_{k \in n} \exp\left(\text{LeakyReLU}\left(\bar{a}_i^T [W_l^T \bar{t}_i || W_l^T \bar{t}_k]\right)\right)} \quad (2)$$

$$\bar{t}_i^{l+1} = \sigma\left(\sum_{j \in n} \alpha_{ij}^l W_l^T \bar{t}_j^l\right) \quad (3)$$

where $l \in \mathbb{R}^{d \times d}$ and $W_l \in \mathbb{R}^{2d}$ are learnable parameters of the l th layer. σ is a scalar indicating the attention score between node i and its neighborhood node j . \bar{t}_j^l represents the feature of node i in the l th layer, with $\bar{t}_0^l = t_i$ initialized from the text representation X^t . As such, we obtain the final textual graph representation $G^t = \{t_1^g, t_2^g, \dots, t_n^g\}$.

Visual Graph Encoder. For an image I , we utilize a pre-trained toolkit developed by [39] to generate a set of regions, denoted as $R = \{r_1, r_2, \dots, r_k\}$. Each region is resized to 224×224 and then divided into p patches. Subsequently, we utilize a pre-trained Vision Transformer (ViT-B/32) [40] along with an additional MLP as the image encoder to encode each region and obtain visual representations. Thus, the representation of the i th region that can be denoted as $V_i = \{v_i^{cls}, v_i^1, v_i^2, \dots, v_i^p\}$, where cls is the representation of [CLS] token and $V_i \in \mathbb{R}^{(p+1) \times d}$. Subsequently, we use the [CLS] token as representation

¹ <https://spacy.io/>

of each region, resulting in the final visual representations $X^v = \{v_1^{cls}, v_2^{cls}, \dots, v_k^{cls}\}$ for the image, where $X^v \in \mathbb{R}^{k \times d}$. For constructing visual graphs, we build edges between each region according to the cosine similarity of representations. We establish an edge between two regions if their cosine similarity exceeds a predefined threshold, denoted as η . Then, we also model the visual graphs with 2-layer graph attention networks. The detailed definition of the graph attention network has been provided in Eq. (2) and Eq. 3, and we omit its explicit description here for brevity. (3). Consequently, we obtain the final visual graph representation $G^v = \{v_1^g, v_2^g, \dots, v_r^g\}$.

Multimodal Graph Fusion Module. In the multimodal graph fusion module, we first concatenate the unimodal graph representations G^v and G^t as $G^{[v,t]}$. Then, we employ M stacked multi-head self-attention layers to fuse the two graph representations. In each layer, the output can be computed as follows:

$$head_i = softmax(\frac{(G^{[v,t]} W_k^i)^T}{\sqrt{d/h}} (G^{[v,t]} W_k^i)) (G^{[v,t]} W_v^i) \quad (4)$$

$$\hat{G} = norm(G^{[v,t]} + MLP([head_1 \parallel head_2 \parallel \dots \parallel head_h])) \quad (5)$$

where $W_q^i \in \mathbb{R}^{d \times \frac{d}{h}}$, $W_k^i \in \mathbb{R}^{d \times \frac{d}{h}}$ and $W_v^i \in \mathbb{R}^{d \times \frac{d}{h}}$ are query, key, and value projection matrices, respectively. "norm" denotes the layer normalization, " \parallel " denotes the concatenation operation and h denotes the number of the head in multi-head self-attention layer. After the processing of M attention layers, we denote the representations for the last attention layer as $G = \{g_1, g_2, \dots, g_{r+n}\}$.

To concisely present the subsequent label-aware graph contrastive learning, we employ the mean-pooling operation followed by normalization operation to perform dimensionality reduction on textual graph representation G^t , visual graph representation G^v and multimodal graph representation G , respectively, thereby obtaining the final graph features \bar{t}^g , \bar{v}^g and \bar{g} .

3.2. Label-aware graph contrastive learning

To enhance the sensitivity of the model to the label-aware distribution and improve its ability to learn graph representations concerning sentiment labels, we propose the Label-aware Graph Contrastive Learning method, as depicted in the bottom right of Fig. 2. To achieve this, we design a label-aware unimodal contrastive loss before the graph fusion stage, followed by a label-aware multimodal contrastive loss after the graph fusion. These loss functions are specifically designed to optimize the model's performance in capturing sentiment-related information within the graph representations.

Label-aware Unimodal Contrastive Loss. Previous Vision-and-Language Pre-training (VLP) methods [32,41] have shown that aligning unimodal features prior to fusion facilitates cross-modal learning. These methods achieve feature alignment by employing image-text contrastive loss, where unimodal features from the same instance are treated as positive samples for each other, while unimodal features from different instances serve as negative samples. Although effective for VLP tasks, these approaches are not directly applicable to multimodal sarcasm detection. This is because in sarcasm detection, samples exhibit correlations based on sentiment labels, unlike in VLP tasks where samples are independent and lack categorical relationships.

To address this challenge, we propose a label-aware unimodal contrastive loss (LUCL) for multimodal sarcasm detection, as it is not reasonable to label unimodal features from all other instances as negative samples. In Fig. 2, visual and textual graphs are labeled as positive samples for each other, indicated by the blue squares along the diagonal. Conversely, instances with different labels from the current instance are labeled as negative samples, represented by the gray squares. In contrast to VLP tasks, where the negative samples commonly include other instances (represented by white squares), considering these instances as negative samples in the context of sarcasm detection would disrupt the negative sample space due to the label correlation among instances.

Consequently, we exclude these contentious samples in negative sample space.

In the following, we briefly introduce the proposed LUCL. Suppose we have a text-image pair $p_i = \{\bar{t}_i^g, \bar{v}_i^g\}$ that belongs to class c , we first define its negative sample space. Specifically, considering a data mini-batch of size b , we regard the other image-text pairs that do not belong to class c as negative samples. Thus, we can obtain the negative sample space for p_i as Ψ_i . Subsequently, we calculate the softmax-normalized visual-to-textual and textual-to-visual similarity in terms of the negative sample space. The definition can be stated as follows:

$$\rho^{v2t} = \frac{\exp(s(\bar{v}_i^g, \bar{t}_i^g)/\tau_u)}{\sum_{k \in \Psi_i} \exp(s(\bar{v}_i^g, \bar{t}_k^g)/\tau_u)} \quad (6)$$

$$\rho^{t2v} = \frac{\exp(s(\bar{t}_i^g, \bar{v}_i^g)/\tau)}{\sum_{k \in \Psi_i} \exp(s(\bar{t}_i^g, \bar{v}_k^g)/\tau)} \quad (7)$$

where $s(\cdot, \cdot)$ is the similarity function, τ_u indicates learning temperature. Let y^{v2t} and y^{t2v} denote the ground-truth one-hot similarity, where negative pairs have a probability of 0 and the positive pair has a probability of 1. The LUCL loss is defined as the cross-entropy between ρ and y :

$$L_{luc1} = \frac{1}{2} [CE(y^{v2t}, \rho^{v2t}) + CE(y^{t2v}, \rho^{t2v})] \quad (8)$$

where CE represents the cross-entropy function.

Label-aware Multimodal Contrastive Loss. To enable the model to learn the label-aware features from multimodal graphs, we use label-aware multimodal contrastive loss during the fusion process in the MGF module. As illustrated in the LMCL in Fig. 2, we divide the multimodal graphs within each batch into positive and negative examples according to their labels. For example, in Fig. 2, for multimodal graphs with sarcasm labels, the graphs in the batch sharing the same sarcasm label indicate positive examples (blue squares), while graphs with non-sarcasm labels are marked as negative examples (gray squares).

In the following, we give the mathematical expressions to better understand the proposed LMCL. For a given batch of size b , we define $f(\bar{g}_i) = \bar{g}_k | \bar{g}_i \in \{\bar{g}_1, \bar{g}_2, \dots, \bar{g}_b\}$ as a set of instance. We introduce loss function L_{lmcl}^{ij} aimed at maximizing the similarity between pairs of sample \bar{g}_i and \bar{g}_j within the same sarcastic class. This loss is formulated as:

$$L_{lmcl}^{ij} = -\beta_{ij} \log \frac{\exp(s(\bar{g}_i, \bar{g}_j)/\tau)}{\sum_{k \in f(\bar{g}_i)} \exp(s(\bar{g}_i, \bar{g}_k)/\tau)} \quad (9)$$

where $s(\cdot, \cdot)$ is the similarity function, τ indicates learning temperature, and β_{ij} denotes the class indicator to ensure that the LMCL loss only applies to samples of the same class. Specifically, if \bar{t}_i and \bar{t}_j share the same sarcastic label, β_{ij} is set as 1, encouraging their closer proximity in the latent space. Conversely, if their labels differ, β_{ij} will be set as 0 to avoid unnecessary clustering. Mathematically, β_{ij} can be defined as follows:

$$\begin{cases} \beta_{ij} = 1, & \text{if } S_i = S_j \\ \beta_{ij} = 0, & \text{else} \end{cases} \quad (10)$$

where S_i and S_j represents the sarcastic labels for \bar{t}_i and \bar{t}_j respectively. Thus, for a given sample \bar{t}_i , we can formalize the overall loss as $L_{lmcl}^i = \sum_{j \in f(\bar{g}_i)} L_{lmcl}^{ij}$. This formulation encourages samples from the same class as \bar{g}_i to cluster together.

3.3. Back-translation data augmentation

Data augmentation has consistently served as a fundamental strategy for enhancing contrastive objectives. In graph contrastive learning, existing methods [34,42] adopt augmentation schemes in the structural domain, such as uniformly dropping edges or shuffling features. However, data augmentation in the structural domain potentially results in the loss of valuable information within the graph, which may hinder

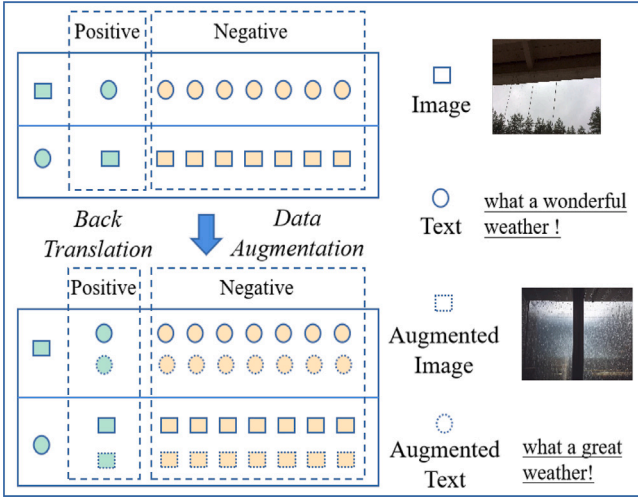


Fig. 3. The number of positive and negative samples in a batch for the proposed LUCI loss, with and without back-translation data augmentation. The green circles and squares denote the text and image data for the current instance, and we treat them as positive samples of each other. The orange circles and squares denote text and image data from other instances that belong to different categories from the current instance, which we treat as negative samples.

the model's ability to capture meaningful patterns [35]. To mitigate this issue, a more effective approach is to perform data augmentation in the data domain [43]. This involves generating a more diverse set of positive and negative samples while preserving the semantic integrity of the original data. To this end, we propose back-translation data augmentation, which generates additional visual and textual data, thereby expanding the positive and negative sample space in the data domain.

For text augmentation, we design the text-based back-translation. This method [44,45] has proven effective in generating diverse paraphrased sentences while preserving the semantic meaning of the original content. Therefore, we employ the back-translation technique to create positive and negative text samples in graph contrastive learning. This process involves utilizing a pre-trained back-translation model to translate an existing text in language E into another language C , and subsequently translating it back to language E , resulting in an augmented text. In contrast to the previous methods [44,45], we employ a large language model for back-translation augmentation, as illustrated in Fig. 2. Specifically, we first define a text instruction “Translate this text into Chinese and then back into English”. Using this instruction, we leverage the GPT-3.5 model [15] to translate existing text $S = \{w_i\}_{i=1}^n$ into Chinese and then translate it back to obtain an augmented text \tilde{S} .

Regarding the image augmentation method, we introduce cross-modal back-translation for data augmentation, as depicted in Fig. 2. The process involves two main components: the BLIP model [16] for caption generation given an image, and the stable-diffusion model [17] for generating an image based on the input textual prompt. Thus, we first employ the pre-trained BLIP model to generate the caption for a given image I , which can be defined as:

$$C = \text{BLIP}(I, \theta_c) \quad (11)$$

where θ_c denotes the pre-trained parameters of the BLIP model [16], C denotes the generated caption. Subsequently, we utilize the caption as a prompt and employ the pre-trained stable-diffusion model [17] to translate the prompt back to an augmented image \tilde{I} :

$$\tilde{I} = \text{Diffusion}(C, \theta_d) \quad (12)$$

where θ_d denotes the pre-trained parameters of the stable-diffusion model. The process of cross-modal back-translation for image augmentation, visualized in Fig. 8, allows for the generation of diverse visual content while preserving the semantics of the original images.

Table 1
Statistics of the experimental data.

Dataset	Label	Train	Val	Test
MMSD	Sarcasm	8642	959	959
	Non-sarcasm	11 174	1451	1450
	All	19816	2410	2409
MMSD2.0	Sarcasm	9572	1042	1037
	Non-sarcasm	10 240	1368	1372
	All	19816	2410	2409

Fig. 3 illustrates the number of positive and negative samples in a batch for the proposed label-aware unimodal contrastive loss (LUCI), with and without back-translation data augmentation. This data augmentation results in a doubling of the scale of positive and negative samples within the data domain compared to the original dataset.

3.4. Training loss

Our model optimizes three losses, including a classification loss and two contrastive losses.

In sarcasm classification, we first use mean pooling to perform dimensionality reduction on multimodal graph representations $G = \{g_1, g_2, \dots, g_{r+n}\}$, and get the final graph representations q . After that, we use the cross-entropy loss function as the sarcasm classification loss:

$$L_{ce} = \text{CrossEntropy}(\text{GELU}(qW_{ce} + b_{ce})) \quad (13)$$

The two contrastive losses can be simply added to the classification loss as the final loss:

$$L = L_{ce} + \lambda_{lucI} L_{lucI} + \lambda_{lmcl} L_{lmcl} \quad (14)$$

where λ_{lucI} and λ_{lmcl} are coefficients to balance the different training losses.

4. Experimental setup

In this section, we first introduce the experimental setup. Then, we present the comparative results and conduct an ablation study with more model variants. Furthermore, we show the advantage of our model by a case study and visualization experiment. Finally, we conduct a qualitative analysis of the proposed back-translation data augmentation.

4.1. Datasets

We demonstrate the effectiveness of our method on two public datasets which are MMSD and MMSD2.0. Both datasets collect data from Twitter, each text-image pair is labeled by a single sentiment. Besides, MMSD2.0 conducts data optimization to address the issues in MMSD by removing the spurious cues and fixing unreasonable annotation, for multi-modal sarcasm detection. For a fair comparison, we follow the experimental settings of prior work [5], which divides the data into training, validation, and test sets in a ratio of 80%:10%:10%. The detailed statistics for the MMSD and MMSD2.0 datasets are listed in Table 1.

4.2. Implementation details

We utilize the pre-trained Roberta model to embed the input text and employ the pre-trained ViT to embed image regions, where both embedding sizes are set to 768. In visual graph modeling, we extract 36 regions from each image, and the regions are established edges between with cosine similarity score over the threshold $\eta = 0.6$. In the multimodal graph fusion module, we set the number of

self-attention layers to 6. In image augmentation, we employ the pre-trained BLIP2 [16] model without further fin-tuning to extract the captions, and use the stable diffusion model [17] pre-trained on LAION-5B dataset to generate augmented images. In text augmentation, we utilize a pre-trained toolkit developed by [15] to generate augmented texts. During the training stage, we utilize Adam as the optimizer with a learning rate of $2e-5$. Besides, we set weight decay as $5e-3$, batch size as 64, and dropout rate as 0.5 to train the model. To avoid overfitting, we apply early stopping with a patience of 5. In label-aware graph contrastive learning, we set both the temperatures τ_u and τ_m to 0.07. Following [5,8,13], we use Accuracy, Precision, Recall, and F1-score to measure the model performance.

4.3. Baseline models

We compare our proposed model LGCL with a series of strong baselines, summarized as follows:

Unimodal Baselines. For text-modality methods, we adopt TextCNN [46], Bi-LSTM [47], SMSD [48] which employs a self-matching network to capture incongruity information for sarcasm detection, and BERT [49] and Roberta [38] are two pre-trained models for text classification. For image-modality methods, we employ the pooled feature of the pre-trained Resnet model and the [CLS] token obtained by the pre-trained ViT model to detect sarcasm.

Multimodal Baselines. For multimodal methods, we consider the following baseline methods for comparison. These include HFM [5], which proposed a hierarchical fusion model for multimodal sarcasm detection. Res-Att [20] directly concatenated visual and textual features for multimodal sarcasm prediction. Att-BERT [20] proposed different attention strategies to detect sarcasm. DIP [50] introduced a channel-wise reweighting strategy to model the uncertain correlation. Additionally, we also evaluate against recent graph-based methods, such as InCrossMGs [6], which employed a heterogeneous graph structure to capture ironic features from different perspectives. CMGCN [13] constructed a cross-modal graph for each instance to explicitly draw the ironic relations between different modalities. HKEmodel [7] proposed a hierarchical framework for sarcasm detection by exploring atomic-level and composition-level congruities based on graph neural networks. MILNet [8] designed three graphs to capture multimodal incongruities. Multi-view CLIP [12] introduced a correction dataset called MMSD2.0, and they also presented a novel framework to leverage multi-grained cues from multiple perspectives. DynRT-Net [21] modeled the dynamic mechanism to restrict the model from dynamically adjusting to diverse image-text pairs.

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. First, our proposed LGCL consistently outperforms existing baselines across all the sarcasm datasets, demonstrating the effectiveness of our model in multimodal sarcasm detection. We also conduct significance tests on our LGCL against the baseline models revealing a significant improvement in terms of most evaluation metrics. Additionally, in comparison to multimodal methods utilizing Bert as the textual backbone network, our model achieves a 1.62% improvement in accuracy over the Multi-view CLIP model on the MMSD dataset. Similarly, when using Roberta as the textual backbone network, our model also outperforms the DynRT-Net model with a 0.52% improvement in Accuracy on the MMSD dataset. Besides, our LGCL model consistently outperforms previous graph-based methods (e.g., HKEmodel, MILNet), indicating the superiority of applying label-aware graph contrastive learning. Moreover, methods relying on the text modality consistently exhibit superior performance compared to

those based solely on the image modality, emphasizing that the core expression of sarcastic and non-sarcastic information predominantly resides within the text modality. Furthermore, methods leveraging both image and text modalities consistently outperform the unimodal baselines across the board, suggesting that harnessing information from both modalities proves more effective for multimodal sarcasm detection. Finally, when considering the MMSD2.0 dataset, the performance of all text and multimodal baselines experiences a varying degree of decrease. This can be attributed to the removal of spurious cues in the text data, which further impacts the final results.

5.2. Ablation study

In this section, we evaluate the performance of the multimodal graph fusion module, label-aware graph contrastive learning, and back-translation data augmentation, as listed in Table 3. The results indicate that our model achieves the best performance when composing all these components. Additionally, the proposed multimodal graph fusion (MGF) module demonstrates competitive performance compared to previous graph-based methods, showing its effectiveness in fusing multimodal graphs. On this foundation, incorporating label-aware unimodal contrastive loss (LUCL) and label-aware multimodal contrastive loss (LMCL) can further improve the model's performance. This demonstrates that LUCL and LMCL can enhance the model's ability to recognize and utilize shared sentiment cues within the graph representations. Moreover, the utilization of back-translation data augmentation to augment positive and negative samples in the data domain proves effective for graph contrastive learning, resulting in improved performance.

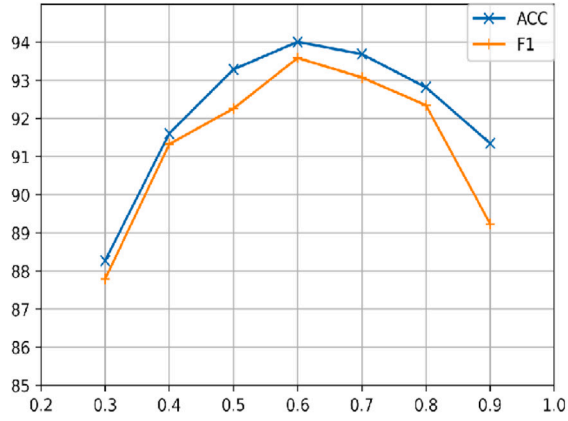
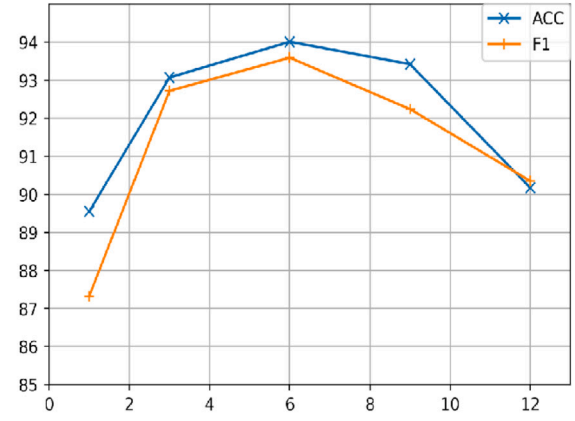
We also analyze the impact of the label-aware strategy in unimodal contrastive learning to further demonstrate the effectiveness of the proposed label-aware unimodal contrastive loss (LUCL). We conducted experiments by adding two different unimodal contrastive losses to the multimodal graph fusion module (MGF). The experimental results are shown in 4. In the table, "UCL" represents the traditional unimodal contrastive loss, which is consistent with the previous Vision-and-Language Pre-training (VLP) methods, treating unimodal features from the same instance as positive samples for each other, while considering all other instances as negative samples. "LUCL" represents the proposed label-aware contrastive loss, which constrains negative sample retrieval to instances with different labels from the target instance. Note that incorporating UCL did not yield a significant improvement in the model's performance, and in some evaluation metrics, there was even a decline. Conversely, equipping LUCL resulted in a noticeable enhancement. From the results, we conclude that restricting the retrieval of negative samples based on sentiment labels effectively prevents the model from considering data from the same label as negative samples, thus avoiding disruption in the negative sample space. And, this further lead to improved performance.

Furthermore, we conducted data augmentations in both the structural and data domains while testing the impacts on graph contrastive learning. Specifically, we employed a uniformly dropping edge scheme in the structural domain, and in the data domain, we utilized the proposed back-translation augmentation scheme. The results are shown in Table 5. Among them, "MGF+LGCL" denotes the model with the proposed label-aware graph contrastive learning (LGCL), "MGF+LGCL+UDeA" denotes the model with label-aware graph contrastive learning and uniformly dropping edge augmentation, and "MGF+LGCL+BTrA" denotes the model with label-aware graph contrastive learning and back-translation augmentation. Obviously, the model based on back-translation augmentation achieves better performance than the model based on uniformly dropping edge augmentation. This demonstrates that data augmentation in the data domain, as opposed to in the structural domain, does not require disrupting the graph structure, thus preserving information in the graph more effectively and resulting in a more significant improvement in graph contrastive learning.

Table 2

Comparison results for sarcasm detection. We use † to indicate the graph-based models. Results with ‡ indicate that the model uses RoBERTa as the textual backbone network, while others use BERT as the textual backbone network.

	Model	MMSD				MMSD2.0			
		ACC (%)	P (%)	R (%)	F1 (%)	ACC (%)	P (%)	R (%)	F1 (%)
text	TextCNN	80.03	74.29	76.39	75.32	71.61	64.62	75.22	69.52
	Bi-LSTM	81.9	76.66	78.42	77.53	72.48	68.02	68.08	68.05
	SMSD	80.9	76.46	75.18	75.82	73.56	68.45	71.55	69.97
	BERT	83.85	78.72	82.27	80.22	74.78	70.52	76.39	73.34
	RoBERTa	87.55	82.09	84.33	83.19	79.66	76.74	75.70	76.21
image	Resnet	64.76	54.41	70.8	61.53	65.50	61.17	54.39	57.58
	Vit	67.83	57.93	70.07	63.4	72.02	65.26	74.83	69.72
text+image	HFM	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
	D&R Net	84.02	77.97	83.42	80.6	–	–	–	–
	Res-BERT	84.80	77.80	84.15	80.85	–	–	–	–
	Att-BERT	86.05	80.87	85.08	82.92	80.03	76.28	77.82	77.04
	InCrossMGs†	86.10	81.38	84.36	82.84	–	–	–	–
	CMGCN†	86.54	–	–	82.73	79.83	75.82	78.01	76.9
	HKEmodel†	87.36	81.84	86.48	84.09	76.5	73.48	71.07	72.25
	Multi-view CLIP	88.33	82.66	88.65	85.55	85.64	80.33	88.24	84.1
	MILNet†‡	89.50	85.16	89.16	87.11	–	–	–	–
	DIP‡	92.97	91.95	94.08	93.01	–	–	–	–
	DynRT-Net‡	93.49	–	–	93.21	–	–	–	–
	LGCL(ours)†	89.95	86.27	90.05	88.12	85.76	85.19	87.04	86.11
	LGCL(ours)†‡	94.01*	92.67*	95.39*	93.59*	–	–	–	–

(a) Analysis of the cosine similarity threshold in visual graph (η)

(b) Analysis of the number of self attention layer (M)

Fig. 4. Hyper-parameter analysis of the LGCL model on the MMSD dataset.**Table 3**

The ablation results of our model.

Model	ACC (%)	Pre (%)	Rec (%)	F1 (%)
MILNet	89.50	85.16	89.16	87.11
MGF	91.59	88.42	92.37	90.35
+LUCL	92.46	89.36	92.94	91.12
+LUCL+BTrA	92.95	89.87	93.41	91.61
+LMCL	92.76	89.52	93.09	91.27
+LMCL+BTrA	93.11	89.95	93.66	91.77
+LUCL+LMCL	93.56	91.37	94.52	92.92
+LUCL+LMCL+BTrA	94.01	92.67	95.39	93.59

Table 4

Performance of using different unimodal contrastive losses.

Model	ACC (%)	Pre (%)	Rec (%)	F1 (%)
MGF	91.59	88.42	92.37	90.35
MGF+UCL	91.41	88.29	92.95	90.56
MGF+LUCL	92.46	89.36	92.94	91.12

On this basis, we also experimented with other text augmentation methods (paraphrasing) to show the effectiveness of the proposed model. The results are shown below. It can be observed that compared

to the GPT3.5-based approach, the data augmented by paraphrasing did not significantly improve the model's performance. Upon close examination of the texts generated by paraphrasing, we noticed that while they generally preserved the overall meaning of the original texts, many ironic features were lost. This issue directly contributed to the poor performance of contrastive learning. However, when we used the GPT3.5-based back-translation augmentation method, due to GPT3.5's strong zero-shot generalization ability, the augmented texts retained the original meaning to a great extent, thus facilitating the improvement of contrastive learning.

Last but not least, we also conduct the experiments by augmenting only one modality in the proposed back-translation data augmentation. Table 7 illustrates the results. We can observe that image data augmentation “+LGCL+BTrA^{image}” yields relatively small benefits to the model, whereas text data augmentation “+LGCL+BTrA^{text}” brings more significant improvements to the model's performance. We attribute it to the fact that image data augmentation utilizes stable-diffusion models, whose image generation capabilities still require further improvement, resulting in noisy generated image data. In contrast, text data augmentation leverages the powerful zero-shot generation capabilities of large language models, achieving more refined and high-quality text data (see Table 6).

Table 5

Performance of the model with different data augmentations.

Model	ACC (%)	Pre (%)	Rec (%)	F1 (%)
MGF+LGCL	93.56	91.37	94.52	92.92
MGF+LGCL+UDeA	93.75	91.63	95.01	93.29
MGF+LGCL+BTrA	94.01	92.67	95.39	93.59

Table 6

Performance of the model with different data augmentation methods on text modality.

Model	ACC (%)	Pre (%)	Rec (%)	F1 (%)
LGCL	93.56	91.37	94.52	92.92
+LGCL+Paraphrasing	93.37	91.29	94.55	92.93
+LGCL+BTrA ^{text}	93.89	92.12	94.67	93.38

Table 7

Performance of the model with data augmentation on one modality.

Model	ACC (%)	Pre (%)	Rec (%)	F1 (%)
LGCL	93.56	91.37	94.52	92.92
+LGCL+BTrA ^{image}	93.77	91.65	94.63	93.12
+LGCL+BTrA ^{text}	93.89	92.12	94.67	93.38

Table 8

Performance of three different types of models equipped with our LGCL for sarcasm detection.

Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
DIP	92.97	91.95	94.08	93.01
DIP(LGCL)	94.05	93.17	95.02	94.09
DynRT-Net	93.49	–	–	93.21
DynRT-Net(LGCL)	94.77	94.66	94.38	94.52

5.3. Parameter analysis

In this section, we conduct experiments to analyze the hyper-parameters for the proposed LGCL model.

Analysis of the cosine similarity threshold in visual graph. We first investigate the parameter analysis for the cosine similarity threshold η in the visual graph encoder. Fig. 4(a) shows that increasing the threshold of the cosine similarity leads to continuous improvement in accuracy and F1 score, but it decreases accuracy when the threshold is greater than 0.6. Therefore, selecting 0.6 for the cosine similarity threshold is optimal since it achieves the best performance.

Analysis of the number of self-attention layers. We measure the model performance on the ACC and F1 score along with a range of the multi-head self-attention layer number M from 1 to 12. We can see in Fig. 4(b), that the ACC score and F1 score increase until reaching a peak point when M equals 6. Our model achieves the best performance at this point. Then, the model performance begins to decrease as M continues to grow. We guess the performance worsens, probably due to the increase of the model parameter, suggesting that adding more multimodal graph fusion layers might not enhance but impede the performance.

Universal discussion. To verify LGCL can boost different types of models, we respectively select the previous SOTA models DIP and DynRT-Net for sarcasm detection to perform the experiments in Table 8. It can be seen that the proposed LGCL improves the performance steadily for the previous SOTA models. These results further validate the universality of our model.

5.4. Case study

To further demonstrate the effectiveness of our model, we provide a case study. We compare the results predicted based on the model with and without(w/ and w/o) label-aware graph contrastive learning (LGCL). As shown in Fig. 5, we can observe that, for these complex

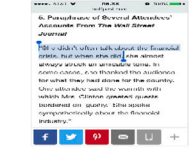

Image	Text	w/ LGCL	w/o LGCL
	.. she is \" the most interesting candidate in the world \" ! # tmcitw # dosequis # hillary # nyprimary # feelthebern	Sarcasm	Non-sarcasm
	happy valentines day ! ! # valentines # valentinesdaycards # valentinstag # cupidday # snrtg # folloforfolloback # follo4follo # sales .	Sarcasm	Non-sarcasm

Fig. 5. Case study with and without label-aware graph contrastive learning(LGCL). The phrases in red font contain strong ironic cues.

cases, it is challenging for the pure multimodal graph fusion module (MGF) to capture the user's sentiments because MGF only captures the atomic-aware graph relations within individual instances. When there are numerous interfering elements in the samples, it can boost the model's ability to capture the correct ironic cues. For example, in the first case, the sentiment of the example is solely determined by the phrase "interesting candidate" in the text, while the presence of other irrelevant words can introduce interference and hinder the model's understanding of the real sentiment for the example. Focusing solely on the atomic-level relations between text and visual graphs within individual instances can lead to incorrect outcomes. In contrast, LGCL relying on label priors to unearth ironic cues from a global perspective of multimodal data, assists the model to detect ironic clues more accurately in complex text.

5.5. Visualization

To verify that the label-aware unimodal contrastive loss (LUCL) and label-aware multimodal Contrastive loss (LMCL) can enhance the model to detect the shared sentiment cues correlated with sentiment labels, we conducted visualization experiments on the MMSD dataset. Firstly, we conduct attention visualization to demonstrate the efficacy of LUCL loss in facilitating precise alignment between the visual and textual graph representations. Besides, we depicted the multimodal graph distribution to illustrate that LMCL loss can cluster multimodal graph representations for instance pairs with the same label, and separate those pairs that have disparate labels.

Attention visualization In Fig. 6, we present the attention visualizations extracted from within the self-attention layers in the multimodal graph fusion module, demonstrating the ability of our label-aware unimodal contrastive loss (LUCL) to perform the unimodal alignment. Particularly, we highlight the regions that are most associated with the specified keywords. The figure reveals that, for the given keywords, our model can pinpoint the relevant regions within the image and allocate augmented attention weights to these regions. This indicates the LUCL's capability to align textual and visual elements at the graph level, thereby facilitating the model's ability to integrate multimodal graphs.

Distribution visualization To visually demonstrate the superiority of label-aware multimodal contrastive loss (LMCL), we visualize the feature distribution on the MMSD val dataset with LMCL loss. Here, we apply the T-SNE algorithm to perform dimensionality reduction for the feature, obtaining a 2-dimensional feature vector distribution visualized in Fig. 7. Fig. 7(a) is the visualization of the distribution generated by the model without LMCL loss, while Fig. 7(b) shows the visualization of the distribution generated by the model with LMCL loss. The figure indicates that by integrating the LMCL loss, an increased



Fig. 6. Attention visualization of some examples.

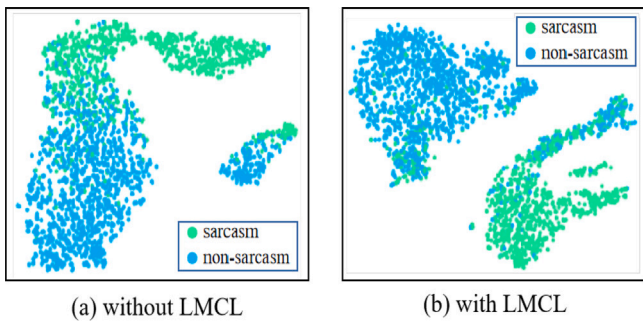


Fig. 7. Cluster visualization of MMSD dataset. Different colored dots represent samples with different labels.

distance between sarcasm and non-sarcasm samples in the graph space leads to a more pronounced degree of data aggregation. This shows that the model distinguishes these data in graph space according to common characteristics inherent to samples sharing the same sentiment.

5.6. Qualitative analysis of data augmentation

We also perform a qualitative analysis of the augmented samples, which are generated by the proposed back-translation data augmentation (BTrA). Concretely, we select three samples and present both their original data forms and the augmented data forms (see Fig. 8). As shown in Fig. 8, we observed that the data augmented through back-translation becomes more diverse while preserving the semantic information consistent with the original data. For example, for the third data in Fig. 8, The enhanced image has essentially restored “the snowfall scene in the wilderness” from the original image. Meanwhile, the enhanced textual semantics also remain largely unchanged. Owing to the semantic in-variance after data augmentation, the graph structures constructed with the original data and the augmented data also maintain consistency in graph semantic space. Therefore, the proposed BTrA can facilitate label-aware graph contrastive learning in the data domain.

6. Conclusion and future work

In this paper, we propose a novel method called label-aware graph contrastive learning to address the limitation of existing multimodal




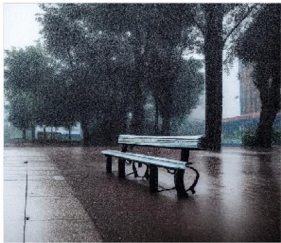


Original data	Augmented data
 <p><user> <user> too bad you don 't get such great cereal in australia huh ... # justacerealbox</p>	 <p><user> <user> Too bad, you haven't eaten such good cereal in Australia, haha... justacerealbox</p>
 <p>The view from my classroom lovely weather</p>	 <p>The scenery from my classroom lovely weather.</p>
 <p>happy spring ! loving all the blossoming flowers happening here ! so beautiful !</p>	 <p>Happy spring! I love all the blossoming flowers happening here! So beautiful!</p>

Fig. 8. Qualitative Analysis of Data Augmentation. We show three samples with their original data forms and the augmented data forms.

sarcasm detection approaches in capturing label-aware connections in the graph semantic space related to sentiment labels. Specifically, we construct unimodal graphs for each instance and fuse them into a graph semantic space to obtain multimodal graphs. Then, we introduce two label-aware graph contrastive losses: the label-aware unimodal contrastive loss and the label-aware multimodal contrastive loss. These losses enhance the model's awareness of shared sentiment cues embedded within the graph representations associated with sentiment labels. To further improve our contrastive learning process, we introduce back-translation augmentation for both textual and visual data, which generates a larger number of positive and negative samples using pre-trained back-translation techniques. Extensive experiments on publicly available benchmark datasets demonstrate the superior performance of our proposed method compared to state-of-the-art baseline approaches. In future work, we plan to further explore the application of contrastive learning in multimodal sarcasm detection and delve deeper into its potential benefits.

CRedit authorship contribution statement

Yiwei Wei: Writing – review & editing, Writing – original draft, Methodology, Funding acquisition. **Maomao Duan:** Writing – review

& editing, Methodology. **Hengyang Zhou**: Visualization, Validation. **Zhiyang Jia**: Software, Project administration. **Zengwei Gao**: Writing – review & editing. **Longbiao Wang**: Writing – review & editing, Writing – original draft.

Declaration of competing interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Data availability

Data will be made available on request.

References

- [1] A. Joshi, P. Bhattacharyya, M.J. Carman, Automatic sarcasm detection: A survey, *ACM Comput. Surv.* 50 (5) (2017) 1–22.
- [2] J. Tepperman, D. Traum, S. Narayanan, “Yeah right”: sarcasm recognition for spoken dialogue systems, in: *Ninth International Conference on Spoken Language Processing*, 2006.
- [3] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, Sarcasm as contrast between a positive sentiment and negative situation, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 704–714.
- [4] O. Tsur, D. Davidov, A. Rappoport, ICWSM—A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews, in: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, (1) 2010, pp. 162–169.
- [5] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in twitter with hierarchical fusion model, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2506–2515.
- [6] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, R. Xu, Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4707–4715.
- [7] H. Liu, W. Wang, H. Li, Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement, 2022, arXiv preprint arXiv:2210.03501.
- [8] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, L. Nie, Mutual-enhanced incongruity learning network for multi-modal sarcasm detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, (8) 2023, pp. 9507–9515.
- [9] Y. Tian, N. Xu, R. Zhang, W. Mao, Dynamic routing transformer network for multimodal sarcasm detection, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2023, pp. 2468–2480.
- [10] R. Schifanella, P. De Juan, J. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in: *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 1136–1145.
- [11] Y. Bin, X. Shang, B. Peng, Y. Ding, T.-S. Chua, Multi-perspective video captioning, in: *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5110–5118.
- [12] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, R. Xu, MMSD2.0: Towards a reliable multi-modal sarcasm detection system, in: *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 10834–10845.
- [13] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, Multi-modal sarcasm detection via cross-modal graph convolutional network, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 1767–1777.
- [14] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.
- [15] OpenAI, GPT-3.5 technical report, 2021, <https://www.openai.com/research/gpt-3.5>.
- [16] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023, arXiv preprint arXiv:2301.12597.
- [17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2022, pp. 10684–10695.
- [18] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, R. Xu, MMSD2.0: Towards a reliable multi-modal sarcasm detection system, in: *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 10834–10845, URL <https://aclanthology.org/2023.findings-acl.689>.
- [19] N. Xu, Z. Zeng, W. Mao, Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3777–3786.
- [20] H. Pan, Z. Lin, P. Fu, Y. Qi, W. Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1383–1392.
- [21] Y. Tian, N. Xu, R. Zhang, W. Mao, Dynamic routing transformer network for multimodal sarcasm detection, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 2468–2480.
- [22] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
- [23] W.L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in: *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.
- [24] Y. Sui, X. Wang, J. Wu, M. Lin, X. He, T.-S. Chua, Causal attention for interpretable and generalizable graph classification, in: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1696–1705.
- [25] X. Kou, B. Luo, H. Hu, Y. Zhang, Nase: Learning knowledge graph embedding for link prediction via neural architecture search, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2089–2092.
- [26] W. Fan, Y. Ma, Q. Li, E. He, J. Zhao, J. Tang, Graph convolutional machine for context-aware recommender systems, in: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM*, 2019, pp. 1059–1066.
- [27] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, W. Xu, Consert: A contrastive framework for self-supervised sentence representation transfer, 2021, arXiv preprint arXiv:2105.11741.
- [28] T. Gao, X. Yao, D. Chen, Simcse: Simple contrastive learning of sentence embeddings, 2021, arXiv preprint arXiv:2104.08821.
- [29] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, H. Ma, Clear: Contrastive learning for sentence representation, 2020, arXiv preprint arXiv:2012.15466.
- [30] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *International Conference on Machine Learning, PMLR*, 2020, pp. 1597–1607.
- [31] X. Chen, K. He, Exploring simple siamese representation learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.
- [32] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International Conference on Machine Learning, PMLR*, 2021, pp. 8748–8763.
- [33] L. Wang, E.-P. Lim, Z. Liu, T. Zhao, Explanation guided contrastive learning for sequential recommendation, in: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 2017–2027.
- [34] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, Y. Shen, Graph contrastive learning with augmentations, in: *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5812–5823.
- [35] Y. Zhu, Y. Xu, F. Yu, Q. Liu, S. Wu, L. Wang, Graph contrastive learning with adaptive augmentation, in: *Proceedings of the Web Conference 2021*, 2021, pp. 2069–2080.
- [36] H. Yang, H. Chen, S. Pan, L. Li, P.S. Yu, G. Xu, Dual space graph contrastive learning, in: *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1238–1247.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
- [39] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. 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*, 2018, pp. 6077–6086.
- [40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
- [41] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, in: *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 9694–9705.
- [42] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, X. Xie, Self-supervised graph learning for recommendation, in: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 726–735.

- [43] E.D. Cubuk, B. Zoph, J. Shlens, Q.V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.
- [44] Q. Xie, Z. Dai, E. Hovy, T. Luong, Q. Le, Unsupervised data augmentation for consistency training, in: *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6256–6268.
- [45] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, 2018, *arXiv preprint* arXiv:1808.09381.
- [46] Y. Chen, *Convolutional Neural Network for Sentence Classification* (Master's thesis), University of Waterloo, 2015.
- [47] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Netw.* 18 (5–6) (2005) 602–610.
- [48] T. Xiong, P. Zhang, H. Zhu, Y. Yang, Sarcasm detection with self-matching networks and low-rank bilinear pooling, in: *The World Wide Web Conference*, 2019, pp. 2115–2124.
- [49] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, *arXiv preprint* arXiv:1810.04805.
- [50] C. Wen, G. Jia, J. Yang, DIP: Dual incongruity perceiving network for sarcasm detection, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, 2023, pp. 2540–2550.