G^{2}SAM: Graph-Based Global Semantic Awareness Method for Multimodal Sarcasm Detection

Yiwei Wei*[1,5], Shaozu Yuan*†[2], Hengyang Zhou [5], Longbiao Wang ‡[1,4],
Zhiling Yan [2], Ruosong Yang [2], Meng Chen [3]

[1] Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] JD AI Research, Beijing, China [3] Yep AI, Melbourne, Australia
[4] Huiyan Technology (Tianjin) Co., Ltd, Tianjin, China
[5] China University of Petroleum(Beijing) at Karamay, Karamay, China

Abstract

Multimodal sarcasm detection, aiming to detect the ironic sentiment within multimodal social data, has gained substantial popularity in both the natural language processing and computer vision communities. Recently, graph-based studies by drawing sentimental relations to detect multimodal sarcasm have made notable advancements. However, they have neglected exploiting graph-based global semantic congruity from existing instances to facilitate the prediction, which ultimately hinders the model's performance. In this paper, we introduce a new inference paradigm that leverages global graph-based semantic awareness to handle this task. Firstly, we construct fine-grained multimodal graphs for each instance and integrate them into the semantic space to draw graph-based relations. During inference, we leverage global semantic congruity to retrieve k-nearest neighbor instances in semantic space as references for voting on the final prediction. To enhance the semantic correlation of representation in semantic space, we also introduce label-aware graph contrastive learning to further improve the performance. Experimental results demonstrate that our model achieves state-of-the-art (SOTA) performance in multimodal sarcasm detection. The code will be available at GGSAM.

Introduction

With the growing reliance on social media platforms like Twitter and Reddit for expressing sentiments, the accurate detection of ironic posts (Tay et al. 2018; Pan et al. 2020; Xu, Zeng, and Mao 2020) and efficient analysis of embedded sentiment (Niu et al. 2016; Xu 2017; Yang et al. 2020; Xu and Mao 2017; Xu, Mao, and Chen 2018) within social media data have garnered significant attention from both academia and industry.

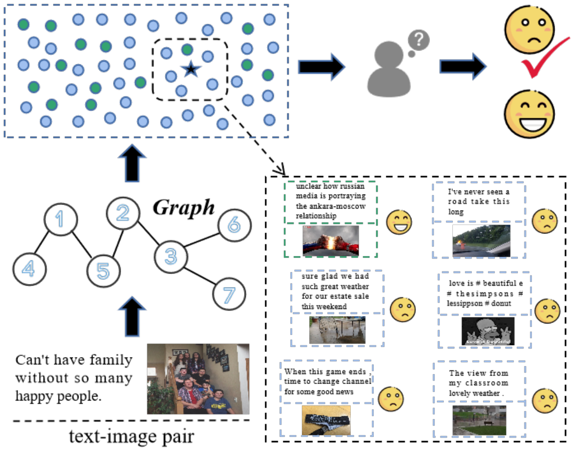

Graph
1
2
3
4
5
6
7
Can't have family without so many happy people.
text-image pair
unclear how russian media is portraying the ankara-moscow relationship
I've never seen a road take this long
sure glad we had such great weather for our estate sale this weekend
love is # beautiful e # thesimpsons # lessippson # donut
When this game ends , time to change channel for some good news
The view from my classroom lovely weather .

Figure 1: Examples of applying global semantic congruity for multimodal sarcasm prediction, where blue dots indicate sarcasm features while the green dots denote non-sarcasm.

Early research primarily focused on textual modalities. Some pattern-based approaches (Davidov, Tsur, and Rappoport 2010; Maynard and Greenwood 2014; Felbo et al. 2017) employed predefined textual patterns to identify specific hashtag labels, leveraging lexical indicators and syntactic rules to identify ironic expressions. To improve the ability to exploit contextual cues, subsequent studies (Tay et al. 2018; Ghosh and Veale 2017; Xiong et al. 2019) explored the sarcasm contexts or the sentiment of sarcasm makers as additional cues to model the congruence level of texts, resulting in consistent improvement. However, relying solely on text modality often fails to fully capture the sentiments of a post.

Compared to purely textual sarcasm detection, multimodal sarcasm detection has been proven more effective for the ever-expanding social media platforms, as they provide supplementary visual cues for sarcasm detection. In recent years, multimodal sarcasm detection has been extensively studied, resulting in considerable progress. Schifanella et al. (2016) adopt attention methods to fuse various modalities. To improve performance, the following methods (Xu, Zeng, and Mao 2020; Liang et al. 2021, 2022; Liu, Wang, and

[*] Both authors contribute equally to this work.
[†] Corresponding authors. Email:{yuanshaozu@jd.com}
[‡] Co-corresponding authors with Shaozu Yuan.

Li 2022; Wei et al. 2023; Qiao et al. 2023; Wen, Jia, and Yang 2023) designed various modules to capture the contradictory relationships between text and image. Considering drawing the intricate sentiment connections between modalities is necessary for sarcasm detection, more recent works, such as InCrossMGs (Liang et al. 2021), CMGCN (Liang et al. 2022), HKEmodel (Liu, Wang, and Li 2022) and MILNet (Qiao et al. 2023), apply graph neural network (GNN) for this task, achieving significant advancements.

Although promising, existing graph-based methods mainly focus on how to draw multimodal graphs and integrate graph features, while neglecting the potential benefits of utilizing global graph-based semantic congruity from existing instances to facilitate prediction. From the view of semantic space, multimodal posts with the same labels typically exhibit more analogous graph representations than other posts (Wang et al. 2017). By incorporating this global semantic congruity into the prediction process, there is potential for further improvement in the model's performance. As illustrated in Figure 1, for instance, the post being predicted shows more semantic congruity with the other sarcasm posts, indicating a higher likelihood of being classified as sarcasm. This global semantic perspective provides a new insight for designing the inference paradigm.

In this paper, we introduce the Graph-Based Global Semantic Awareness Method (G^{2}SAM), a new inference paradigm that leverages global semantic congruity for multimodal sarcasm detection. Concretely, we propose a Fine-grained Graph-aligned (FGM) model, a simple yet effective framework to align and fuse fine-grained unimodal graphs into a graph-based global space to capture contradictory sentiment cues. During the prediction stage, we utilize the graph-based semantic congruity to select the k-nearest neighbor instances of the case to be predicted in the semantic space and vote for the final result, as depicted in Figure 1. Since the graph-based representations in the semantic space can be produced in advance, this process only incurs minimal computational cost. In this stage, it is crucial to ensure the semantic congruity between the retrieved k instances and the case to be predicted, as the k-nearest neighbor instances directly determine the final prediction. However, the fine-grained graph-aligned model lacks direct awareness of the inference process, leading to weakened semantic correlation of graph-based representations and harming the prediction performance. To alleviate this problem, we introduce Label-aware Graph Contrastive Learning (LGCL), which constrains graph-based representations with the same labels to be more similar in the semantic space. This enhances the semantic correlation of retrieved k-nearest neighbor instances, increasing the likelihood of finding k instances with the same label as the case to be predicted and improving the model's performance. To our best knowledge, this is the first work to apply global semantic congruity with label-aware graph contrastive learning for multimodal classification. The experiment shows G^{2}SAM achieves state-of-the-art on the public multimodal sarcasm detection dataset (Cai, Cai, and Wan 2019).

To summarize, the main contributions of this paper are two-fold:

- This paper introduces a novel inference paradigm to multimodal sarcasm detection by applying graph-based global semantic awarenes (G^{2}SAM). It achieves the SOTA performance in multimodal sarcasm detection.
- To the best of our knowledge, this is the first work to exploit global semantic congruity combined with label-aware contrastive learning for multimodal classification, thus opening up new possibilities for the application of global semantic awareness in related research fields.

Related Work

Multimodal sarcasm detection. Multimodal sarcasm detection has emerged as a more challenging problem with the increasing requirement of analyzing multimodal posts on social media. Schifanella et al. (Schifanella et al. 2016) was the first to tackle this problem as a multimodal classification task, by concatenating manually designed multimodal features. Besides, HFM (Cai, Cai, and Wan 2019) proposed a hierarchical fusion model that fuses features extracted from textual and visual modalities, using a new multimodal sarcasm detection dataset based on Twitter. D&R Net (Xu, Zeng, and Mao 2020) constructed the Decomposition and Relation Network to represent the contextual contrast and capture the semantic association between multimodal information, while Att-BERT (Pan et al. 2020) applied the co-attention and self-attention to learn both intra-modality and inter-modality congruity information. In terms of graph-based methods, InCrossMGs (Liang et al. 2021) explored the sentiment inconsistencies by constructing in-modal and cross-modal graphs, whereas CMGCN (Liang et al. 2022) constructed a cross-modal graph for each entity to draw ironic relations between textual and visual information. Furthermore, HKEmodel (Liu, Wang, and Li 2022) modeled hierarchical congruity based on cross-attention mechanism and graph neural networks. And, MILNet (Qiao et al. 2023) built three graphs to learn the local and global incongruities. However, these methods fail to exploit graph-based semantic congruity from existing instances to guide the model toward making better predictions during inference.

Graph Neural Networks. Graph Neural Networks (GNNs) have significantly advanced the learning representations of graph-structured data in recent years. The concept was first proposed by GCN (Kipf and Welling 2016), which applied a spectral graph convolution operation to the input data. This operation essentially involves multiplying the input features by the graph Laplacian matrix. Since then, there have been numerous extensions and variations of GNNs, such as graph attention networks (Veličković et al. 2017) that applied self-attention to weight the importance of neighboring nodes, and GraphSAGE (Hamilton, Ying, and Leskovec 2017), which aggregated information from a node's local neighborhood using a fixed-size feature vector. More recent works have explored the applications of GNNs for various areas, such as graph classification (Sui et al. 2022), link prediction (Kou et al. 2020), and recommendation systems (Fan et al. 2019). Despite their successes, the potential of GNNs for multimodal modeling tasks can be further explored from the global semantic perspective.

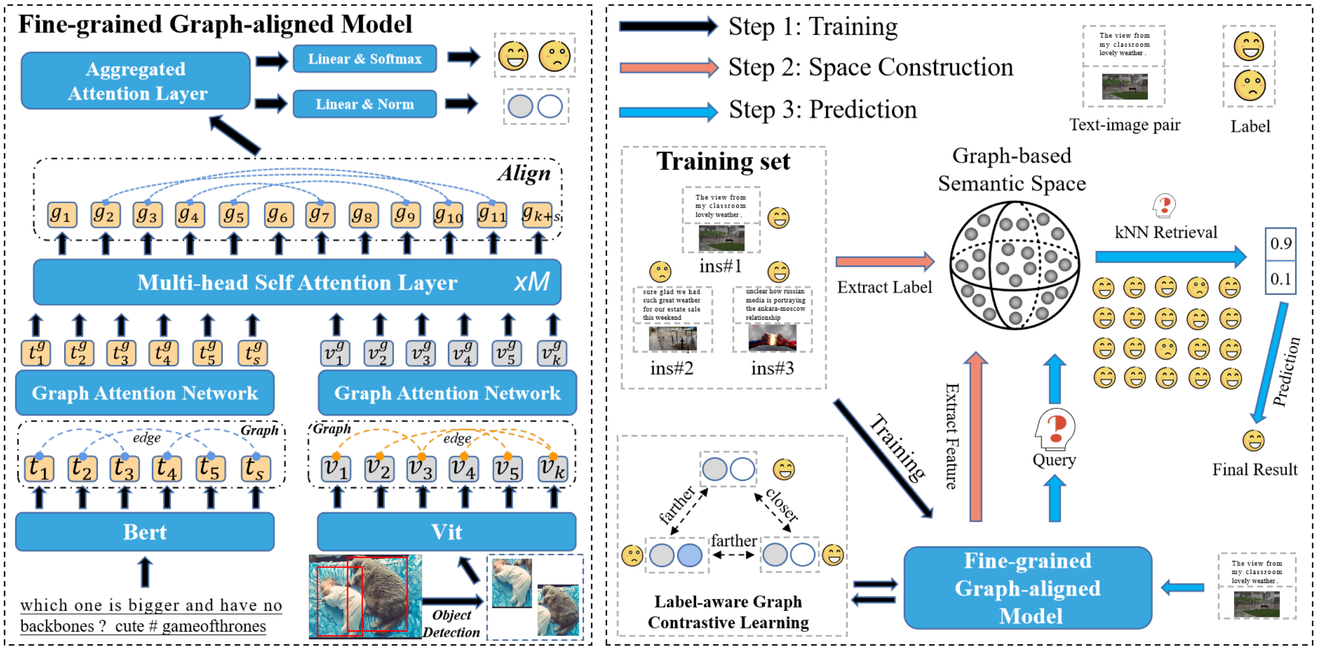

Fine-grained Graph-aligned Model
Aggregated Attention Layer
Linear & Softmax
Linear & Norm
Align
g_1 g_2 g_3 g_4 g_5 g_6 g_7 g_8 g_9 g_{10} g_{11} g_{k+s}
Multi-head Self Attention Layer
xM
t_1^g t_2^g t_3^g t_4^g t_5^g t_s^g
v_1^g v_2^g v_3^g v_4^g v_5^g v_k^g
Graph Attention Network
Graph Attention Network
edge
Graph
Graph
edge
t_1 t_2 t_3 t_4 t_5 t_s
v_1 v_2 v_3 v_4 v_5 v_k
Bert
Vit
which one is bigger and have no backbones ? cute # gameofthrones
Object Detection
Step 1: Training
Step 2: Space Construction
Step 3: Prediction
The view from my classroom lovely weather .
Text-image pair
Label
Training set
The view from my classroom lovely weather .
ins#1
sure glad we had such great weather for our estate sale this weekend
unclear how russian media is portraying the ankara-moscow relationship
ins#2
ins#3
Extract Label
Graph-based Semantic Space
kNN Retrieval
0.9
0.1
Prediction
Final Result
Extract Feature
Query
Training
farther
closer
farther
Label-aware Graph Contrastive Learning
Fine-grained Graph-aligned Model
The view from my classroom lovely weather .

Figure 2: The overall architecture of our model. The left figure presents fine-grained graph-aligned model to generate multi-modal graph-based representation. And the right figure illustrates how to generate the result with graph-based global semantic congruity, where this inference process can be aware via label-aware graph contrastive learning during the training stage.

Methodology

As illustrated in Figure 2, the architecture of G^{2}SAM is mainly composed of three procedures: 1) to construct graph representations for both textual and visual modalities and fuse the graphs, we proposed a fine-grained graph-aligned model (FGM); 2) we project the graph-based representation to semantic space, and k-nearest neighbor instances features are selected for semantic prediction according to global semantic incongruity; 3) to ensure the semantic correlation of retrieved k-nearest neighbor instances and improve the prediction performance, we also introduce label-aware graph contrastive learning.

Fine-grained Graph-aligned Model

To achieve accurate multimodal sarcasm detection, it is essential to recognize contradictory sentimental cues from different modalities. Therefore, building fine-grained relationships between text and images is necessary to better capture these sentimental differences. For this purpose, we first draw the fine-grained graphs: token-level graphs for the text and region-level graphs for the image. We then design a graph-aligned fusion module to align and fuse fine-grained graph in the semantic space.

Fine-grained Feature Extraction. Given an input text-image pair (T, I), the first step is to extract textual and visual representations. For text T, the pre-trained BERT-base (Devlin et al. 2018) is applied to transform T into a sequence of token-level feature $X^t = \{t_1, t_2, ..., t_s\}$, where $X^t \in \mathbb{R}^{s \times d}$. For image I, we utilize a pre-trained toolkit (Anderson et al. 2018) to extract k regions, denoted as $R = \{r_1, r_2, ..., r_k\}$. And each region is resized to 224 × 224 and divided into p patches. Subsequently, a pre-trained Vision Transformer ViT-B/32 (Dosovitskiy et al. 2020) is employed as the image encoder to capture the semantic feature for each region. Thus, the representation of the i-th region can be denoted as $I_i = \{v_i^{cls}, v_i^1, v_i^2, ..., v_i^p\}$, where cls is the representation of [CLS] token and $I_i \in \mathbb{R}^{[p+1] \times d}$. We use the representation of [CLS] token to represent each region, resulting in the final region-level feature $X^v = \{v_1^{cls}, v_2^{cls}, ..., v_k^{cls}\}$ for the image, where $X^v \in \mathbb{R}^{k \times d}$.

Fine-grained Graph Modeling. To measure the relations for sentimental cues of each modality, we utilize the obtained fine-grained representations to construct the fine-grained unimodal graphs. To construct the textual graphs, we follow HKEmodel (Liu, Wang, and Li 2022), where the text tokens serve as graph nodes and relations extracted from spaCy[1] between words denote as edges. For the visual graphs, we build edges between each object region according to the cosine similarity score of representation. Then, we model the graphs in textual and visual modalities with 2-layer graph attention networks (GAT) (Veličković et al. 2017). As such, we obtain the token-level textual graph $G^t = \{t_1^g, t_2^g, ..., t_s^g\}$ and region-level visual graph $G^v = \{v_1^g, v_2^g, ..., v_k^g\}$ respectively.

Graph-aligned Fusion Module. In contrast to previous complicated fusion methods (Liang et al. 2022; Liu, Wang, and Li 2022; Qiao et al. 2023), we design a relatively simple approach to align and fuse the graphs of text and images. Concretely, we first concatenate the visual and textual graph representations G^v and G^t as $G^{[v,t]}$. Then, we employ M

[1] https://spacy.io/

stacked self-attention layers to align and fuse the two graph representations. In each layer, the output can be computed as follows:

$$G = softmax(\frac{(G^{[v,t]}W_q)^T}{\sqrt{d}}(G^{[v,t]}W_k))(G^{[v,t]}W_v) \quad (1)$$

where $W_q \in \mathbb{R}^{d \times d}$, $W_k \in \mathbb{R}^{d \times d}$ and $W_v \in \mathbb{R}^{d \times d}$ are query, key, and value projection matrices, respectively. For simplicity, we omit the residual connection and layer normalization for each self-attention layer. Here, we denote the representations for the last attention layer as $G^M = \{g_1, g_2, ..., g_{k+s}\}$ and $[,]$ represents the concatenation operation.

Considering the representations of graph-aligned fusion can not be directly applied for classification, we employ an aggregated attention layer to perform dimensionality reduction for sarcasm classification, which is formalized as:

$$\tilde{r}_i = GELU(g_i W_1 + b_1)W_2 + b_2 \quad (2)$$

$$\tilde{q} = \sum_{i=1}^{k+s} exp(\frac{\tilde{r_i}}{\sum_{j=1}^{k+s} \tilde{r}_j})(g_i) \quad (3)$$

$$q = GELU(\tilde{q}W_3 + b_3) \quad (4)$$

where $GELU$ is the activation function, $q \in \mathbb{R}^d$.

Graph Semantic Congruity Prediction

During inference, G^{2}SAM utilizes graph-based global semantic congruity from existing instances to make predictions, as depicted in Figure 2 (right). To achieve this, we first project sentimental graphs for each instance into the semantic space. Subsequently, we retrieve the k-nearest neighbors in the semantic space, which are then utilized for prediction.

Graph-based semantic space. To better illustrate this process, we simplify the fine-grained graph-aligned model that maps the graph to graph-based semantic representation as $\phi(\cdot)$. For each instance (x_i, y_i) from the training set, we use $q_i = \phi(x_i)$ map x_i to graph-based representation q_i. Thus, the graph-based semantic space $\acute{D}$ can be constructed by a single forward pass over each training instance: $\acute{D} = (q_i, y_i)_{i=1}^{N}$, where N is the number of instances in the graph-based semantic space. Note that this process can be calculated in advance without any additional training because the graph-based representations in the semantic space are only utilized for prediction.

Prediction. To determine the final prediction $\hat{y}_{Cr} \in [0, 1]$ for the current case x during test time, we select k-nearest neighbor (kNN) instances according to global semantic congruity and voted for the final prediction. Concretely, the graph-based representation $\phi(x)$ for the input text-image pair x acts as a query q to retrieve the k-nearest neighbor graph representations $\mathcal{N} = \{(q_i, y_i)\}_{i=1}^{k}$ from graph-based semantic space. Here, we apply Euclidean distance to measure the global graph-based semantic congruity between q and $\mathcal{N}$, for the magnitude of vectors is more appropriate to reflect the semantic correlations in the semantic space. And this prediction process $\hat{y}_{kNN}$ can be defined as:

$$\hat{y}_{kNN} = \sum_{i=1}^{k} \alpha_i y_i, \ a_i = \frac{e^{-\|q_i - q\|_2^2/\tau}}{\sum_j e^{-\|q_j - q\|_2^2/\tau}} \quad (5)$$

where $\| \ \|_2^2$ indicates the euclidean distance, τ is the kNN temperature, and α_i denotes the weight of the i-th neighbor. According to semantic correlation, the nearest neighbor instances are more likely to have the same label as most retrieved cases. Thus, we apply a voting mechanism to obtain the final prediction $\hat{y}$ as follows:

$$\begin{cases} \hat{y} = 1, \text{ if } \hat{y}_{kNN} \geq 0.5 \\ \hat{y} = 0, \text{ if } \hat{y}_{kNN} < 0.5 \end{cases} \quad (6)$$

Label-aware Graph Contrastive Learning

As mentioned above, the k-retrieved nearest neighbor (kNN) instances directly determine the final sarcasm detection in G^{2}SAM. However, the fine-grained graph-aligned model lacks direct awareness of the inference process. This can diminish the semantic correlation of the graph-based representation, thereby negatively impacting the prediction performance. To solve this issue, we propose label-aware graph contrastive learning (LGCL) to enhance graph-based semantic correlation in semantic space. In this principle, the graph-based feature with the same label is forced to be in semantic congruity in semantic space, which ensures the k-retrieved instances are more likely to have the same label as the predicting case. The key to utilizing contrastive learning is how to construct positive or negative examples. Previous work (You et al. 2020) selects one as a positive example in a mini-batch, while all other samples serve as negatives. However, this is not adapted to G^{2}SAM for two reasons: 1) due to the absence of positive instances in a mini-batch, it requires the design of complex data augmentation methods to generate contrastive pairs, which introduces additional computational costs. 2) selecting one instance as a positive example is not reasonable as there may be multiple instances in a mini-batch with the same positive label.

To handle the above two problems, we introduce label-aware graph contrastive learning, where the instances in the mini-batch are directly treated as positive or negative examples according to their labels. Specifically, the sarcasm instances in the batch are labeled as positive, while the non-sarcasm instances are labeled as negative examples. To provide a detailed explanation of our graph contrastive learning algorithm, we present its step-by-step process in Algorithm 1. In the algorithm, $Norm$ signifies the normalization function, $gather$ denotes gathering values along with an index, and $\hat{\tau}$ represents the temperature of the graph contrastive learning. Especially, there are two steps in the algorithm: the first step is to generate the unmask label L_t according to the sarcasm labels in the batch; in the second step, we compute the similarity matrix l and leverage the unmask label L_t and the similarity matrix l to calculate the contrastive loss L_{LGCL} to optimize the model.

Training loss

We optimize the multimodal graph fusion model by minimizing the cross entropy loss L_{ce} as previous work (Cai, Cai, and Wan 2019), which can be defined as:

$$L_{ce} = CrossEntropy(GELU(qW_{ce} + b_{ce})) \quad (7)$$

Algorithm 1: LGCL Algorithm

```
Input: The label in the batch is L, which is a list of all samples, assuming that the samples are divided into two categories: sarcasm (1), non-sarcasm (0); The fine-grained graph-aligned model $\phi()$; the text-image pairs x; C denotes the length of $\mathrm{L_c}$; $S$ denotes length of L.
Output: Label-aware graph contrastive loss $\mathrm{L_{LGCL}}$
initialize $\mathrm{L_c} = [\mathrm{L} - 0, \mathrm{L} - 1]$ and $\mathrm{L_t} = \mathrm{list}()$
for $\mathrm{i} = 1; \mathrm{i} \leq \mathrm{C}; \mathrm{i} + +$ do
    initialize $\acute{\mathrm{L}}_\mathrm{t}$ = list()
    for $\mathrm{j} = 1; \mathrm{j} \leq \mathrm{S}; \mathrm{j} + +$ do
        if $\mathrm{L_c[i][j]}$ equals 0 then
            $\acute{\mathrm{L}}_\mathrm{t}.\mathrm{append(j)}$
        end
    end
    $\mathrm{L_t}.\mathrm{append}(\acute{\mathrm{L}}_\mathrm{t})$
end
$\acute{\mathrm{q}} = \mathrm{Norm}(\phi(\mathrm{x})), \acute{\mathrm{I}} = \acute{\mathrm{q}}@\acute{\mathrm{q}}^\mathrm{T}$
$\mathrm{l} = \mathrm{LogSoftmax}(\acute{\mathrm{I}})/\hat{\tau}.\mathrm{view}(-1)$
$\mathrm{L_{cl}} = \mathrm{L_t[L[1]]}$
for $\mathrm{k} = 2; \mathrm{k} \leq \mathrm{S}; \mathrm{k} + +$ do
    $\mathrm{L_{cl}} = \mathrm{concat}(\mathrm{L_{cl}}, \mathrm{L_t[L[k]]} + \mathrm{k} \times \mathrm{S})$
end
$\mathrm{L_{LGCL}} = \mathrm{gather}(\mathrm{l}, \mathrm{index} = \mathrm{L_{cl}})/\mathrm{L_{cl}}.\mathrm{size}(0)$
Return $\mathrm{L_{LGCL}}$
```

Dataset	Label	Train	Val	Test
HFM	Positive	8642	959	959
	Negative	11174	1451	1450
	All	19816	2410	2409

Table 1: Statistics of the experimental data.

Besides, we apply graph contrastive loss L_{LGCL} defined in Algorithm 1 to distinguish graph features. Totally, the training loss can be defined as:

$$L = L_{ce} + \gamma L_{LGCL} \tag{8}$$

where γ is a hyper-parameter to balance different losses.

Experiments

Datasets

The primary experiments were carried out using the publicly accessible multimodal sarcasm detection dataset (Cai, Cai, and Wan 2019). In this dataset, tweets that express sarcasm are considered positive examples, while those that do not express sarcasm are deemed negative examples. Each example in the dataset comprises a text and an associated image. The statistics for the dataset is listed in Table 1.

Experimental Settings

To ensure fairness, we follow previous works (Cai, Cai, and Wan 2019; Liang et al. 2021) for dataset pre-processing. We use pre-trained BERT-base and ViT models for text and image embeddings respectively, both set to size 768. In visual graph modeling, we extract 36 regions per image, creating edges between regions with cosine similarity over 0.6. The graph-aligned fusion module has 6 self-attention layers. During training, we use Adam optimizer with a learning rate of 2e-5, weight decay of 5e-3, batch size of 64, and dropout rate of 0.5. Early stopping with a patience of 5 is applied to prevent overfitting. For graph contrastive learning, the temperature $\hat{\tau}$ is set at 0.07. Performance is measured using Accuracy, Precision, Recall, and F1-score, following Cai, Cai, and Wan (2019); Liang et al. (2022). Macro-average scores are reported to account for imbalanced data distribution.

Baseline Models

To fully validate the performance of G^{2}SAM, we select both unimodal and multimodal baselines.

Unimodal Baselines. For text-modality methods, we utilize TextCNN (Chen 2015), Bi-LSTM (Graves and Schmidhuber 2005), and BERT(Devlin et al. 2018), which is a pre-trained model specifically designed for text classification. As for image-modality methods, we leverage the pooled feature of the pre-trained Resnet model, along with the [CLS] token obtained from the pre-trained ViT model, to detect sarcasm.

Multimodal Baselines. For multimodal methods, we consider the following baseline methods for comparison. These include HFM (Cai, Cai, and Wan 2019), which proposed a hierarchical fusion model for multimodal sarcasm detection. Att-BERT (Pan et al. 2020) proposed different attention strategies to detect sarcasm. DIP (Wen, Jia, and Yang 2023) introduced a channel-wise reweighting strategy to model the uncertain correlation. Additionally, we also evaluate against recent graph-based methods, such as InCrossMGs (Liang et al. 2021), which employed a heterogeneous graph structure to capture ironic features from different perspectives. CMGCN (Liang et al. 2022) constructed a cross-modal graph for each instance to explicitly draw the ironic relations between different modalities. HKEmodel (Liu, Wang, and Li 2022) proposed a hierarchical framework for sarcasm detection by exploring atomic-level and composition-level congruities based on graph neural networks. And MILNet (Qiao et al. 2023) designed three graphs to capture multimodal incongruities.

Experimental Results

Main Results

To evaluate the performance of G^{2}SAM, we summarize the experimental results of various models for the multimodal sarcasm detection task in Table 2. From these results, we can derive several conclusions. Firstly, it is obvious that the models based on the text modality are more competitive against the baselines on image modality, due to the lower information density of the image modality compared to the text modality, which has also been discussed by Hu et al. (2022); Liu, Wang, and Li (2022). Furthermore, the multimodal models outperform the unimodal sarcasm models as they provide a greater number of sentimental cues for sarcasm detection. Generally, our G^{2}SAM achieves the best

MODALITY	METHOD	Acc(%)	Pre(%)	Rec(%)	F1(%)	Macro-average		
						Pre(%)	Rec(%)	F1(%)
image	Resnet	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.68	71.35	68.40
text	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
image+text	HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	Res-BERT	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	InCrossMGs*	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	CMGCN*	86.54	-	-	82.73	-	-	-
	HKEmodel*	87.36	81.84	86.48	84.09	-	-	-
	MILNet*†	88.72	84.97	87.79	86.37	87.75	88.29	88.04
	DIP	89.59	87.76	86.58	87.17	88.46	89.13	89.01
	Ours*	**90.48**	**87.95**	**89.02**	**88.48**	**89.44**	**89.79**	**89.65**

Table 2: Experimental results for sarcasm detection. We use $*$ to indicate the graph-based models. † indicates the reproduced results by unifying the textual backbone with the previous works. As mentioned by Liang et al. (2022), p-value $\geq$ 0.05 indicates a significant improvement for this task.

Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
MILNet	88.72	84.97	87.79	86.37
FGM	89.01	85.73	87.15	86.43
FGM+kNN	89.86	86.82	87.89	87.35
FGM+LGCL	89.33	86.29	87.56	86.92
FGM+kNN+LGCL	**90.48**	**87.95**	**89.02**	**88.48**

Table 3: The ablation results of our model. To show the superiority of FGM, we also provide the result of the previous SOTA graph-based model MILNet for comparison.

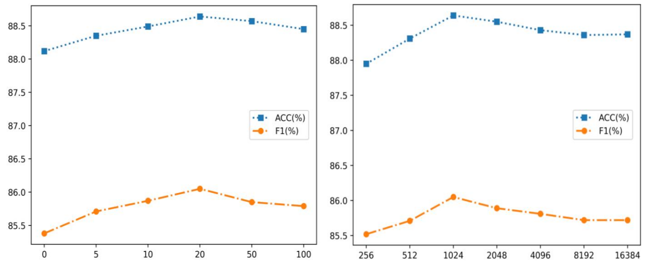
Model	ACC(%)	Pre(%)	Rec(%)	F1(%)
ViT	67.25	57.48	69.93	63.10
G^{2}SAM(ViT)	68.75	58.20	70.59	63.79
BERT	84.74	79.27	83.52	81.34
G^{2}SAM(BERT)	85.65	80.09	84.31	82.15
DIP	89.59	87.76	86.58	87.17
G^{2}SAM(DIP)	90.21	88.01	87.45	87.73

Table 4: Performance of three different types of models equipped with our G^{2}SAM for sarcasm detection.

performance across all metrics, demonstrating the advantage of exploring graph-based semantic awareness. Compared to the previous SOTA model DIP, G^{2}SAM exceeds it by 0.89% and 1.31% in terms of accuracy and F1. Given that the margin of improvement for most previous models is less than 1%, this constitutes a significant enhancement, demonstrating the strong performance of the G^{2}SAM.

Ablation Study

In order to evaluate the effectiveness of different components, we conduct an ablation study in Table 3. Interestingly, despite its simplicity compared to the SOTA graph-based model MILNet, the fine-grained graph-aligned model (FGM) delivers comparable performance, indicating its ability for aligning different unimodal graphs and identifying conflicting sentiment cues across different modalities. In addition, we notice that the improvements can be further improved by leveraging semantic congruity (kNN) for prediction during test time, which proved the superiority of this inference paradigm. Moreover, incorporating label-aware graph contrastive learning (LGCL) into FGM can boost the predicting performance, as LGCL enhances the semantic correlation of representations in the semantic space. Since each module has its own unique strengths, it is evident that the FGM, when equipped with LGCL in the new inference paradigm (kNN), ultimately produces the most superior result.



(a) Performance of the number(k) of neighbors (b) Performance of the size(N) of semantic space

Figure 3: The curve of performance for multimodal sarcasm detection with different settings.

In addition, we also conduct the experiments to verify G^{2}SAM can boost different types of models. We respectively select text-modality, image-modality, and multimodal models for sarcasm detection to perform the experiments in Table 4. It can be seen that the proposed G^{2}SAM improves the performance steadily for the pure text-modality model (BERT). Although the visual modality contains less sentimental information for sarcastic clues, G^{2}SAM still brings some improvement to the image-modality model (ViT). Even though the previous SOTA model, DIP (Qiao et al. 2023), exhibits significant performance, it can still be fur-

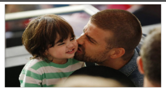
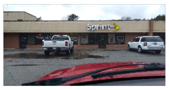
Text-Image Pairs	kNN Prediction						GT
	Instance	label	α	Instance	label	α	
Can't have family without so many happy people.	ins#1	1	0.205	ins#6	1	0.069	Sarcasm 1
	ins#2	1	0.102	ins#7	1	0.065	
	ins#3	1	0.087	ins#8	1	0.065	
	ins#4	1	0.082	ins#9	0	0.063	
	ins#5	1	0.072	ins#10	1	0.057	
 Happy fourth birthday to one of the cutest and happiest kids.	ins#1	0	0.163	ins#6	1	0.068	Non-sarcasm 0
	ins#2	0	0.125	ins#7	0	0.063	
	ins#3	0	0.107	ins#8	0	0.059	
	ins#4	0	0.075	ins#9	0	0.051	
	ins#5	0	0.073	ins#10	0	0.051	
 This is a man who obviously knows how to make good decisions.	ins#1	1	0.092	ins#6	1	0.071	Sarcasm 1
	ins#2	1	0.087	ins#7	0	0.071	
	ins#3	0	0.085	ins#8	1	0.068	
	ins#4	1	0.079	ins#9	1	0.065	
	ins#5	1	0.075	ins#10	0	0.064	

Figure 4: User study for sampled instances. Here, we provide the top 10 retrieved kNN instances for analysis. And α_i denotes the weight of the $i-th$ neighbor.

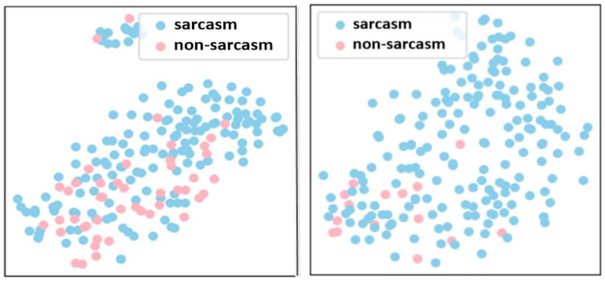
ther enhanced when equipped with G^{2}SAM.

Optimal Settings Exploring

In this section, we conduct experiments to determine the optimal settings for G^{2}SAM. We first investigate the optimal setting for graph semantic incongruity prediction. Figure 3 illustrates the curve for the number of neighbors and instances required to construct a graph-based semantic space. Figure 3(a) shows that increasing the number of neighbor instances leads to continuous improvement in accuracy and F1 score, but it decreases accuracy when k is greater than 20. Therefore, selecting 20 nearest neighbor instances for prediction is optimal since it achieves the best performance. Figure 3(b) demonstrates that the model achieves its best performance when there are 1024 graph-based instances in the semantic space. However, increasing N will harm performance because super abundant instances in the graph space make it difficult for the model to capture similar cases.

Case Study

In Figure 4, we provide a case study to provide a detailed analysis of inference with global semantic congruity. As depicted in the figure, the majority of retrieved reference instances clearly indicate the true label, particularly in the first two cases where sarcasm or non-sarcasm cues are explicitly conveyed. Consequently, these two cases are relatively simple for the model to confidently output the correct answer. However, when the case obscurely conveys sarcasm or non-sarcasm cues, it introduces some difficulty for the model to identify the true label. In the third case, the selected instances are divided into two categories, indicating the case's difficulty in distinguishing the true label. Despite this, we find the vast majority of reference instances still reveal the correct label, showing the ability of G^{2}SAM. This interesting phenomenon aligns with our intuitive understanding that the challenging case tends to have lower classification confidence scores.



(a) without LGCL (b) with LGCL

Figure 5: Distribution of the retrieved top 200 nearest neighbors instances for a sarcasm case. Here, the LGCL is removed in Figure (a).

Visualization

To further visually demonstrate the effectiveness of our LGCL, we visualize the feature distribution around a sarcasm instance. We retrieve the top 200 nearest neighbor instances as reference instances for a sarcasm case, and we employ T-SNE[3] algorithm for dimensionality reduction, obtaining a 2-dimensional feature vector distribution visualized in Figure 5. Figure 5(a) depicts the distribution without LGCL, while Figure 5(b) represents the distribution when the model is equipped with LGCL. From figure 5(b), we find that most retrieved instances exhibit a higher likelihood of indicating the true label (sarcasm), whereas the instances in Figure (a) contain more noise (non-sarcasm). This suggests that the LGCL enhances the semantic correlation of retrieved k-nearest neighbor instances, thereby improving the prediction performance.

Conclusion

In this paper, we propose a novel paradigm for handling multimodal sarcasm detection task by using graph-based global semantic awareness. Specifically, we propose an effective fine-grained graph-aligned model to capture contradictory sentimental cues, and then project the aligned features into semantic space. During the inference stage, we leverage the graph-based semantic congruity to retrieve the k-nearest neighbor instances in the semantic space and make predictions based on them. To improve the performance, we introduce label-aware graph contrastive learning to enhance semantic congruity for graph-based semantic representation. Extensive experiments are conducted to verify the effectiveness of the proposed method. Finally, this work also shows the universality of multimodal sentiment detection tasks, indicating huge potential for extension to other multimodal classification research areas. Due to space limitations, we omitted the relevant discussion in the camera-ready version. We are looking forward to researchers exploring the potential of applying this paradigm to other multimodal domains.

[3] https://github.com/mxl1990/tsne-pytorch

Acknowledgments

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.62176182).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2506–2515.

Chen, Y. 2015. *Convolutional neural network for sentence classification*. Master's thesis, University of Waterloo.

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon (pp. 15–16). *Retrieved from Association for Computational Linguistics website: https://www. aclweb. org/anthology/W10-2914. pdf.*

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, W.; Ma, Y.; Li, Q.; He, E.; Zhao, J.; and Tang, J. 2019. Graph convolutional machine for context-aware recommender systems. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 1059–1066.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*.

Ghosh, A.; and Veale, T. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 482–491.

Graves, A.; and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, 18(5-6): 602–610.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, 1024–1034.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. *arXiv preprint arXiv:2211.11256*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kou, X.; Luo, B.; Hu, H.; and Zhang, Y. 2020. Nase: Learning knowledge graph embedding for link prediction via neural architecture search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2089–2092.

Liang, B.; Lou, C.; Li, X.; Gui, L.; Yang, M.; and Xu, R. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, 4707–4715.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1767–1777.

Liu, H.; Wang, W.; and Li, H. 2022. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. *arXiv preprint arXiv:2210.03501*.

Maynard, D. G.; and Greenwood, M. A. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Lrec 2014 proceedings*. ELRA.

Niu, T.; Zhu, S.; Pang, L.; and El Saddik, A. 2016. Sentiment Analysis on Multi-View Social Data. In Tian, Q.; Sebe, N.; Qi, G.-J.; Huet, B.; Hong, R.; and Liu, X., eds., *MultiMedia Modeling*, 15–27.

Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; and Wang, W. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1383–1392.

Qiao, Y.; Jing, L.; Song, X.; Chen, X.; Zhu, L.; and Nie, L. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9507–9515.

Schifanella, R.; De Juan, P.; Tetreault, J.; and Cao, L. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, 1136–1145.

Sui, Y.; Wang, X.; Wu, J.; Lin, M.; He, X.; and Chua, T.-S. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1696–1705.

Tay, Y.; Luu, A. T.; Hui, S. C.; and Su, J. 2018. Reasoning with Sarcasm by Reading In-Between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1010–1020.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, Y.; Zhang, L.; Deng, H.; Lu, J.; Huang, H.; Zhang, L.; Liu, J.; Tang, H.; and Xing, X. 2017. Learning a discriminative distance metric with label consistency for scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8): 4427–4440.

Wei, Y.; Yuan, S.; Yang, R.; Shen, L.; Li, Z.; Wang, L.; and Chen, M. 2023. Tackling Modality Heterogeneity with Multi-View Calibration Network for Multimodal Sentiment Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5240–5252.

Wen, C.; Jia, G.; and Yang, J. 2023. DIP: Dual Incongruity Perceiving Network for Sarcasm Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2540–2550.

Xiong, T.; Zhang, P.; Zhu, H.; and Yang, Y. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The world wide web conference*, 2115–2124.

Xu, N. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE international conference on intelligence and security informatics (ISI)*, 152–154. IEEE.

Xu, N.; and Mao, W. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2399–2402.

Xu, N.; Mao, W.; and Chen, G. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 929–932.

Xu, N.; Zeng, Z.; and Mao, W. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 3777–3786.

Yang, X.; Feng, S.; Wang, D.; and Zhang, Y. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23: 4014–4026.

You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; and Shen, Y. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823.