Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement

Hui Liu[1]   **Wenya Wang[2,3]**   **Haoliang Li[1]**
[1]City University of Hong Kong
[2]Nanyang Technological University
[3]University of Washington
liuhui3-c@my.cityu.edu.hk, wangwy@ntu.edu.sg, haoliang.li@cityu.edu.hk

Abstract

Sarcasm is a linguistic phenomenon indicating a discrepancy between literal meanings and implied intentions. Due to its sophisticated nature, it is usually challenging to be detected from the text itself. As a result, multi-modal sarcasm detection has received more attention in both academia and industries. However, most existing techniques only modeled the atomic-level inconsistencies between the text input and its accompanying image, ignoring more complex compositions for both modalities. Moreover, they neglected the rich information contained in external knowledge, e.g., image captions. In this paper, we propose a novel hierarchical framework for sarcasm detection by exploring both the atomic-level congruity based on multi-head cross attention mechanism and the composition-level congruity based on graph neural networks, where a post with low congruity can be identified as sarcasm. In addition, we exploit the effect of various knowledge resources for sarcasm detection. Evaluation results on a public multi-modal sarcasm detection dataset based on Twitter demonstrate the superiority of our proposed model.

1 Introduction

Sarcasm refers to satire or ironical statements where the literal meaning of words is contrary to the authentic intention of the speaker to insult someone or humorously criticize something. Sarcasm detection has received considerable critical attention because sarcasm utterances are ubiquitous in today's social media platforms like Twitter and Reddit. However, it is a challenging task to distinguish sarcastic posts to date in light of their highly figurative nature and intricate linguistic synonymy (Pan et al., 2020; Tay et al., 2018).

Early sarcasm detection methods mainly relied on fixed textual patterns, e.g., lexical indicators, syntactic rules, specific hashtag labels and emoji occurrences (Davidov et al., 2010; Maynard and

Figure 1: An example of sarcasm along with the corresponding image and different types of external knowledge extracted from the image. The sarcasm sentence represents the need for some good news. However, the image of the TV program is switched to bad news depicting severe storms (bad weather) which contradicts the sentence.

Greenwood, 2014; Felbo et al., 2017), which usually had poor performances and generalization abilities by failing to exploit contextual information. To resolve this issue, (Tay et al., 2018; Joshi et al., 2015; Ghosh and Veale, 2017; Xiong et al., 2019) considered sarcasm contexts or the sentiments of sarcasm makers as useful clues to model congruity level within texts to gain consistent improvement. However, purely text-modality-based sarcasm detection methods may fail to discriminate certain sarcastic utterances as shown in Figure 1. In this case, it is hard to identify the actual sentiment of the text in the absence of the image forecasting severe weather. As text-image pairs are commonly observed in the current social platform, multi-modal methods become more effective for sarcasm prediction by capturing congruity information between textual and visual modalities (Pan et al., 2020; Xu et al., 2020a; Schifanella et al., 2016; Liu et al., 2021; Liang et al., 2021; Cai et al., 2019).

However, most of the existing multi-modal techniques only considered the congruity level between each token and image-patch (Xu et al., 2020a; Tay et al., 2018) and ignored the importance of multi-granularity (e.g., granularity such as objects, and relations between objects) alignments, which have been proved to be effective in other related tasks, such as cross-modal retrieval (Li et al., 2021b) and

image-sentence matching (Xu et al., 2020b; Liu et al., 2020). In fact, the hierarchical structures of both texts and images advocate for composition-level modeling besides single tokens or image patches (Socher et al., 2014). By exploring compositional semantics for sarcasm detection, it helps to identify more complex inconsistencies, e.g., inconsistency between a pair of related entities and a group of image patches.

Moreover, as figurativeness and subtlety inherent in sarcasm utterances may bring a negative impact to sarcasm detection, some works (Li et al., 2021a; Veale and Hao, 2010) found that the identification of sarcasm also relies on the external knowledge of the world beyond the input texts and images as new contextual information. What's more, it has drawn increasing research interest in how to incorporate knowledge to boost many machine learning algorithms such as recommendation system (Sun et al., 2021) and relation extraction (Sun et al., 2022). Indeed, several studies extracted image attributes (Cai et al., 2019) or adjective-noun pairs (ANPs) (Xu et al., 2020a) from images as visual semantic information to bridge the gap between texts and images. However, constrained by limited training data, such external knowledge may not be sufficient or accurate to represent the images (as shown in Figure 1) which may bring negative effects for sarcasm detection. Therefore, how to choose and leverage external knowledge for sarcasm detection is also worth being investigated.

To tackle the limitations mentioned above, in this work, we propose a novel hierarchical framework for sarcasm detection. Specifically, our proposed method takes both atomic-level congruity between independent image objects and tokens, as well as composition-level congruity considering object relations and semantic dependencies to promote multi-modal sarcasm identification. To obtain atomic-level congruity, we first adopt the multi-head cross attention mechanism (Vaswani et al., 2017) to project features from different modalities into the same space and then compute a similarity score for each token-object pair via inner products. Next, we obtain composition-level congruity based on the output features of both textual modality and visual modality acquired in the previous step. Concretely, we construct textual graphs and visual graphs using semantic dependencies among words and spatial dependencies among regions of objects, respectively, to capture composition-level feature

for each modality using graph attention networks (Veličković et al., 2018). Our model concatenates both atomic-level and composition-level congruity features where semantic mismatches between the texts and images in different levels are jointly considered. Specially, we elaborate the terminology used in our paper again: congruity represents the semantic consistency between image and text. If the meaning of the image and text pair is contradictory, this pair will get less congruity. Atomic is between token and image patch, and compositional is between a group of tokens (phrase) and a group of patches (visual object).

Last but not the least, we propose to adopt the pre-trained transferable foundation models (e.g., CLIP (Radford et al., 2021, 2019)) to extract text information from the visual modality as external knowledge to assist sarcasm detection. The rationality of applying transferable foundation models is due to their effectiveness on a comprehensive set of tasks (e.g., descriptive and objective caption generation task) based on the zero-shot setting. As such, the extracted text contains ample information of the image which can be used to construct additional discriminative features for sarcasm detection. Similar to the original textual input, the generated external knowledge also contains hierarchical information for sarcasm detection which can be consistently incorporated into our proposed framework to compute multi-granularity congruity against the original text input.

The main contributions of this paper are summarized as follows: 1) To the best of our knowledge, we are the first to exploit hierarchical semantic interactions between textual and visual modalities to jointly model the atomic-level and composition-level congruities for sarcasm detection; 2) We propose a novel kind of external knowledge for sarcasm detection by using the pre-trained foundation model to generate image captions which can be naturally adopted as the input of our proposed framework; 3) We conduct extensive experiments on a publicly available multi-modal sarcasm detection benchmark dataset showing the superiority of our method over state-of-the-art methods with additional improvement using external knowledge.

2 Related Work

2.1 Multi-modality Sarcasm Detection

With the rapid growth of multi-modality posts on modern social media, detecting sarcasm for text

and image modalities has increased research attention. Schifanella et al. (2016) first defined multi-modal sarcasm detection task. Cai et al. (2019) created a multi-modal sarcasm detection dataset based on Twitter and proposed a powerful baseline fusing features extracted from both modalities. Xu et al. (2020a) modeled both cross-modality contrast and semantic associations by constructing the Decomposition and Relation Network to capture commonalities and discrepancies between images and texts. Pan et al. (2020) and Liang et al. (2021) modeled intra-modality and inter-modality incongruities utilizing transformers (Vaswani et al., 2017) and graph neural networks, respectively. However, these works neglect the important associations played by hierarchical or multi-level cross-modality dismatches. To address this limitation, we propose to capture multi-level associations between modalities by cross attentions and graph neural networks to identify sarcasm in this work.

2.2 Knowledge Enhanced Sarcasm Detection

Li et al. (2021a) and Veale and Hao (2010) pointed out that commonsense knowledge is crucial for sarcasm detection. For multi-modal based sarcasm detection, Cai et al. (2019) proposed to predict five attributes for each image based on the pre-trained ResNet model (He et al., 2016) as the third modality for sarcasm detection. In a similar fashion, Xu et al. (2020a) extracted adjective-noun pairs(ANPs) from every image to reason discrepancies between texts and ANPs. In addition, as some samples can contain text information for the images, Pan et al. (2020) and Liang et al. (2021) proposed to apply the Optical Character Recognition (OCR) to acquire texts on the images. More recently, Liang et al. (2022) proposed to incorporate objection detection framework and label information of detected visual objects to mitigate modality gap. However, the knowledge extracted from these methods is either not expressive enough to convey the information of the images or is only restricted to a fixed set, e.g., nearly one thousand classes for image attributes or ANPs. Moreover, it should be noted that not every sarcasm post has text on images. To this end, in this paper, we propose to generate a descriptive caption with rich semantic information for each image based on the pre-trained Clipcap model (Mokady et al., 2021), which uses the CLIP (Radford et al., 2021) encoding as a prefix to the caption by employing a simple mapping network

and then fine-tunes GPT-2 (Radford et al., 2019) to generate the image captions.

3 Methodology

Our proposed framework contains four main components: Feature Extraction, Atomic-Level Cross-Modal Congruity, Composition-Level Cross-Modal Congruity and Knowledge Enhancement. Given an input text-image pair, the feature extraction module aims to generate text features and image features via a pre-trained text encoder and an image encoder, respectively. These features will then be fed as input to the atomic-level cross-modal congruity module to obtain congruity scores via a multi-head cross attention model (MCA). To produce composition-level congruity scores, we construct a textual graph and a visual graph and adopt graph attention networks (GAT) to exploit complex compositions of different tokens as well as image objects. The input features to the GAT are taken from the output of the atomic-level module. Due to the page limitation, we place our illustration figure in Figure 6. Our model is flexible to incorporate external knowledge as a "virtual" modality, which could be used to generate complementary features analogous to the image modality for congruity score computation.

3.1 Task Definition & Motivation

Multi-modal sarcasm detection aims to identify whether a given text associated with an image has a sarcastic meaning. Formally, given a multi-modal text-image pair (X_T, X_I), where X_T corresponds to a textual tweet and X_I is the corresponding image, the goal is to produce an output label $y \in \{0, 1\}$, where 1 indicates a sarcastic tweet and 0 otherwise. The goal of our model is to learn a hierarchical multi-modal sarcasm detection model (by taking both atomic-level and composition-level congruity into consideration) based on the input of textual modality, image modality and the external knowledge if chosen.

The reason to use composition-level modeling is to cope with the complex structures inherent in two modalities. For example, as shown in Figure 2, the semantic meaning of the sentence depends on composing *your life*, *awesome* and *pretend* to reflect a negative position, which could be reflected via the dependency graph. The composed representation for text could then be compared with the image modality for more accurate alignment detection.
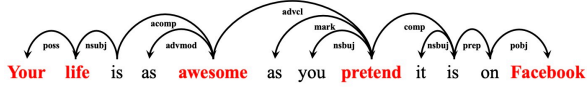
Figure 2: Semantic dependency between words in sarcasm text

3.2 Feature Extraction

Given an input text-image pair (X_T, X_I), where $X_T = \{w_1, w_2, \ldots, w_n\}$ consists of n tokens, we utilize the pre-trained BERT model (Devlin et al., 2019) with an additional multi-layer perceptron (MLP) to produce a feature representation for each token, denoted as $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n]$, where $\mathbf{T} \in \mathbb{R}^{n \times d}$. As for image processing, given the image X_I with the size $L_h \times L_w$, following existing methods (Xu et al., 2020a; Cai et al., 2019; Liang et al., 2021; Pan et al., 2020), we first resize the image to size 224×224. Then we divide each image into r patches and reshape these patches into a sequence, denoted as $\{p_1, p_2, \ldots, p_r\}$, in the same way as tokens in the text domain. Next, we feed the sequence of r image patches into an image encoder to get a visual representation for each patch. Specifically, in this paper, we choose two kinds of image encoders including the pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2020) and a ResNet model (He et al., 2016), both of which are trained for image classification on ImageNet. Hence, the embedding of image patches derived by ViT or ResNet contains rich image label information. Here we adopt the features before the final classification layer to initialize the embeddings for visual modality. We further use a two-layer MLP to obtain the feature representations for $\{p_1, p_2, \ldots, p_r\}$ as $\mathbf{I} = [\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_r]$, where $\mathbf{I} \in \mathbb{R}^{r \times d}$.

3.3 Atomic-Level Congruity Modeling

To measure atomic-level congruity between a text sequence and an image, an intuitive solution is to compute a similarity score between each token and a visual patch directly. However, due to the huge gap between two different modalities, we propose to use cross attention mechanisms with h heads to firstly align the two modalities in the same space, which can be computed as

$$\mathbf{head}_i = \text{softmax}\left(\frac{(\mathbf{TW}_q^i)^\top}{\sqrt{d/h}}(\mathbf{IW}_k^i)\right)(\mathbf{IW}_v^i), \quad (1)$$

where $\mathbf{I} \in \mathbb{R}^{r \times d}$ and $\mathbf{T} \in \mathbb{R}^{n \times d}$ are feature representations of the given text and image, respectively.

$\mathbf{W}_q^i \in \mathbb{R}^{d \times \frac{d}{h}}$, $\mathbf{W}_k^i \in \mathbb{R}^{d \times \frac{d}{h}}$ and $\mathbf{W}_v^i \in \mathbb{R}^{d \times \frac{d}{h}}$ are query, key and value projection matrices, respectively, for $\mathbf{head}_i \in \mathbb{R}^{n \times \frac{d}{h}}$. It is worth noting that we also consider taking image as query, text as key and value for Equation (1). However, we empirically find that the performance is not desired in this case. We conjecture the reason to be the fact that the visual modality may not contain sufficient information and is less expressive compared to the textual modality to provide attentive guidance, which can lead to negative impact of the final performance.

Then, by concatenating all heads followed by a two-layer MLP and a residual connection, we obtain updated text representations $\tilde{\mathbf{T}} \in \mathbb{R}^{n \times d}$ after aligning with the visual modality as

$$\tilde{\mathbf{T}} = \text{norm}(\mathbf{T} + \text{MLP}([\mathbf{head}_1 \| \mathbf{head}_2 \| \ldots \| \mathbf{head}_h])), \quad (2)$$

where "norm" denotes the layer normalization operation and "$\|$" denotes the concatenation operation. Next, to perform atomic-level cross-modal congruity detection, we adopt the inner product as $\mathbf{Q}_a = \frac{1}{\sqrt{d}}(\tilde{\mathbf{T}}\mathbf{I}^\top)$ where $\mathbf{Q}_a \in \mathbb{R}^{n \times r}$ is the matrix consisting of $\mathbf{Q}_a[i, j]$ for i-th row and j-th column representing the similarity score between the i-th token of the text and the j-th patch of the image. Intuitively, different words can have different influence on the sarcasm detection task. For example, noun, verb and adjacent words are usually more important for understanding sarcastic utterances. As such, we feed features of words to a fully-connected (FC) layer with a softmax activation function to model the token importance for sarcasm detection. The final atomic level congruity score $\mathbf{s}_a$ can be obtained by a weighted sum of $\mathbf{Q}_a$ with the importance score of each token as

$$\mathbf{s}_a = \text{softmax}(\tilde{\mathbf{T}}\mathbf{W}_a + \mathbf{b}_a)^\top \mathbf{Q}_a, \quad (3)$$

where $\mathbf{W}_a \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_a \in \mathbb{R}^n$ are trainable parameters in the FC layer for token importance score computation. $\mathbf{s}_a \in \mathbb{R}^r$ contains the predicted atomic-level congruity score corresponding to each of the r patches.

3.4 Composition-Level Congruity Modeling

The composition-level congruity detection considers the more complex structure of both the text and image modalities, compared to the atomic-level computations. To achieve that, we propose to first construct a corresponding textual graph and a visual graph for the input text-image pair. For the

textual graph, we consider tokens in the input text as graph nodes and use dependency relations between words extracted by spaCy[1] as edges, which have been proved to be effective for various graph-related tasks (Liu et al., 2020; Liang et al., 2021). Concretely, if there exists a dependency relation between two words, there will be an edge between them in the textual graph. For the visual graph, given r image patches, we take each patch as a graph node and connect adjacent nodes according to their geometrical adjacency. Additionally, both two kinds of graphs are undirected and contain self-loops for expressiveness.

Then, we model the graphs in text and visual modalities with graph attention networks (GAT) (Veličković et al., 2018). GAT leverages self-attention layers to weigh the extent of information propagated from corresponding nodes. By using GAT, atomic-level semantic information will propagate along with the graph edge to learn composition-level representations for both textual modality and image modality. Here, we take the textual graph for illustration given as

$$\alpha_{i,j}^l = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{v}_l^\top [\boldsymbol{\Theta}_l \mathbf{t}_i^l \| \boldsymbol{\Theta}_l \mathbf{t}_j^l]\right)\right)}{\sum_k \exp\left(\text{LeakyReLU}\left(\mathbf{v}_l^\top [\boldsymbol{\Theta}_l \mathbf{t}_i^l \| \boldsymbol{\Theta}_l \mathbf{t}_k^l]\right)\right)}, \quad (4)$$

$$\mathbf{t}_i^{l+1} = \alpha_{i,i}^l \boldsymbol{\Theta}_l \mathbf{t}_i^l + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \boldsymbol{\Theta}_l \mathbf{t}_j^l, \quad (5)$$

where $k \in \mathcal{N}(i) \cup \{i\}$, $\boldsymbol{\Theta}_l \in \mathbb{R}^{d \times d}$ and $\mathbf{v}_l \in \mathbb{R}^{2d}$ are learnable parameters of the l-th textual GAT layer. $\alpha_{i,j}^l$ is a scalar indicating the attention score between node i and its neighborhood node j. $\mathbf{t}_i^l$ represents the feature of node i in the l-th layer, with $\mathbf{t}_i^0 = \tilde{\mathbf{t}}_i$ initialized from the atomic-level features $\tilde{\mathbf{T}}$. We use $\hat{\mathbf{T}} = [\mathbf{t}_1^{L_T}, \mathbf{t}_2^{L_T}, \ldots, \mathbf{t}_r^{L_T}]$ with $\hat{\mathbf{T}} \in \mathbb{R}^{n \times d}$ to represent the composition-level embeddings of the textual modality after L_T GAT layers that incorporate complex dependencies among related tokens. In some cases, we may not be able to construct a reliable textual graph due to the lack of sufficient words in a sentence or errors from the parser. Hence, we further propose to concatenate $\hat{\mathbf{T}}$ with a sentence embedding $\mathbf{c} \in \mathbb{R}^d$ which is computed by a weighted sum of each word embedding in $\tilde{\mathbf{T}}$:

$$\mathbf{c} = \text{softmax}(\mathbf{T}\mathbf{W}_c + \mathbf{b}_c)^\top \tilde{\mathbf{T}}, \quad (6)$$

with learnable $\mathbf{W}_c \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_c \in \mathbb{R}^n$.

Likewise, we can obtain $\hat{\mathbf{I}} = [\mathbf{i}_1^{L_I}, \mathbf{i}_2^{L_I}, \ldots, \mathbf{i}_n^{L_I}]$, $\hat{\mathbf{I}} \in \mathbb{R}^{r \times d}$ as the composition-level representations in the visual modality. At last, we compute

composition-level alignment scores $\mathbf{s}_p$ between $\hat{\mathbf{T}}$ and $\hat{\mathbf{I}}$ in a similar way as atomic-level congruity as

$$\mathbf{s}_p = \text{softmax}([\hat{\mathbf{T}} \| \mathbf{c}]\mathbf{W}_p + \mathbf{b}_p)^\top \mathbf{Q}_p, \quad (7)$$

where $\mathbf{Q}_p = \frac{1}{\sqrt{d}}([\hat{\mathbf{T}} \| \mathbf{c}]\hat{\mathbf{I}}^\top) \in \mathbb{R}^{(n+1) \times r}$ is the matrix of composition-level congruity between textual modality and visual modality, $\mathbf{W}_p \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_p \in \mathbb{R}^{n+1}$ are trainable parameters. $\mathbf{s}_p \in \mathbb{R}^r$ contains the final predicted composition-level congruity score for each of the r image patches.

3.5 Knowledge Enhancement

While using text-image pair can benefit sarcasm detection compared with only using a single modality , recent works have shown that it might be still challenging to detect sarcasm solely from a text-image pair (Li et al., 2021a; Veale and Hao, 2010). To this end, we explore the effect of fusing various external knowledge extracted from an image for sarcasm detection. For example, the knowledge could be image attributes (Cai et al., 2019), ANPs (Xu et al., 2020a) as they provide more information on the key concepts delivered by the image. However, such information lacks coherency and semantic integrity to describe an image and may introduce unexpected noise, as indicated in Figure 1. To address this limitation, we propose to generate image captions as the external knowledge to assist sarcasm detection. We further compare the effect of each knowledge form in the experiments.

To fuse external knowledge into our model, we treat knowledge X_K as another "virtual" modality besides texts and images. Then the augmented input to the model becomes (X_T, X_I, X_K). As the knowledge is given in textual form, we follow the process of generating text representations to attain the knowledge features. Specifically, we first obtain the input knowledge representations as $\mathbf{K} = [\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_m]$ using BERT with a MLP, which is analogous to $\mathbf{T}$. Then, we propose to reason the congruity score between text and knowledge modalities at atomic-level by following the procedure of computing atomic-level congruity score between text and image modalities (as shown in Equations (1)-(3)) with another set of parameters. Concretely, for cross-modality attentions between texts and knowledge, we replace $\mathbf{I}$ in Equation (1) with $\mathbf{K}$ and $\mathbf{T}$ in Equation (1) with $\tilde{\mathbf{T}}$, which is the updated text representations after aligning with the visual modality. Inheriting information from the image modality, using $\tilde{\mathbf{T}}$ as the query to attend to knowledge enhances deeper interactions

across all the three modalities. By further replacing $\mathbf{T}$ in Equation (2) with $\tilde{\mathbf{T}}$, we denote the atomic-level text representations after aligning with the knowledge by $\tilde{\mathbf{T}}^k$. The similarity matrix between texts and knowledge becomes $\mathbf{Q}_a^k = \frac{1}{\sqrt{d}}(\tilde{\mathbf{T}}^k \mathbf{K}^\top)$. Then the atomic-level congruity score, denoted as $\mathbf{s}_a^k \in \mathbb{R}^m$, can be obtained as

$$\mathbf{s}_a^k = \mathrm{softmax}(\tilde{\mathbf{T}}^k \mathbf{W}_a^k + \mathbf{b}_a^k)^\top \mathbf{Q}_a^k. \tag{8}$$

By adopting the dependency graph for X_K, we further generate the updated knowledge representations $\hat{\mathbf{K}}$ via GAT and obtain the composition-level congruity score $\mathbf{s}_p^k \in \mathbb{R}^m$ between text and knowledge modalities, following the same procedure for text-image composition-level congruity score as described in Section 3.4.

3.6 Training & Inference

Given both the atomic-level and composition-level congruity scores $\mathbf{s}_a$ and $\mathbf{s}_p$, respectively, the final prediction could be produced considering the importance of each image patch for sarcasm detection.

$$\mathbf{p}_v = \mathrm{softmax}(\mathbf{IW}_v + \mathbf{b}_v), \tag{9}$$
$$\mathbf{y}' = \mathrm{softmax}(\mathbf{W}_y[\mathbf{p}_v \odot \mathbf{s}_a \parallel \mathbf{p}_v \odot \mathbf{s}_p] + \mathbf{b}_y), \tag{10}$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times 1}$, $\mathbf{b}_v \in \mathbb{R}^r$, $\mathbf{W}_y \in \mathbb{R}^{2 \times 2r}$ and $\mathbf{b}_y \in \mathbb{R}^2$ are trainable parameters, $\mathbf{p}_v \in \mathbb{R}^r$ is a r-dim attention vector, $\odot$ is element-wise vector product. It is flexible to further incorporate external knowledge by reformulating Equation (10) to

$$\mathbf{y}' = \mathrm{softmax}(\mathbf{W}_y^k[\mathbf{p}_v \odot \mathbf{s}_a \parallel \mathbf{p}_v \odot \mathbf{s}_p \parallel$$
$$\mathbf{p}_k \odot \mathbf{s}_a^k \parallel \mathbf{p}_k \odot \mathbf{s}_p^k] + \mathbf{b}_v^k),$$

where $\mathbf{s}_a^k$ and $\mathbf{s}_p^k$ are atomic-level and composition-level congruity scores between post and external knowledge. $\mathbf{p}_k \in \mathbb{R}^m$ measures the importance of each word in the knowledge obtained by $\mathbf{p}_k = \mathrm{softmax}(\mathbf{KW}_v^k + \mathbf{b}_v^k)$. The entire model can be trained in an end-to-end fashion by minimizing the cross-entropy loss given the ground-truth label y.

4 Experiments

4.1 Dataset

Table 1: Statistics of the dataset

	Training	Development	Testing
Sarcasm	8642	959	959
Non-Sarcasm	11174	1451	1450
All	19816	2410	2409
Token Length	16.91	16.92	17.13
Entity	3.76	3.71	3.84

We evaluate our model on a publicly available multi-modal sarcasm detection dataset in English constructed by Cai et al. (2019). The statistics are shown in Tabel 1. Based on our preliminary analysis, the average numbers of tokens and entities in a text are approximately 17 and 4, respectively, where complex compositions among atomic units have a higher chance of being involved. This finding provides the basis for our framework using atomic-level and composition-level information to capture hierarchical cross-modality semantic congruity.

4.2 Implementation

For a fair comparison, following the pre-processing in (Cai et al., 2019; Liang et al., 2021; Xu et al., 2020a), we remove samples containing words that frequently co-occur with sarcastic utterances (e.g., *sarcasm*, *sarcastic*, *irony* and *ironic*) to avoid introducing external information. The dependencies among tokens are extracted using spaCy toolkit. For image preprocessing, we resize the image to 224×224 and divide it into 32×32 patches (i.e., $p = 7$, $r = 49$). For knowledge extraction, we extract image attributes following (Cai et al., 2019), ANPs following (Xu et al., 2020a) and image captions via Clipcap (Mokady et al., 2021).

Next, we employ a pre-trained BERT-base-uncased model[2] as textual backbone network to obtain initial embeddings for texts and knowledge, and choose the pre-trained ResNet and ViT[3] modules as visual backbone networks to extract initial embeddings for images. These textual and visual representations are mapped to 200-dim vectors by corresponding MLPs. We use Adam optimizer to train the model. The dropout and early-stopping are adopted to avoid overfitting. The details of implementations are listed in Table 6 in Appendix. Our code is avaliable at `https://github.com/less-and-less-bugs/HKEmodel`.

4.3 Baseline Models

We divide the baseline models into three categories: text-modality methods, image-modality methods and multi-modality methods. For text-based models, we adopt **TextCNN** (Kim, 2014), **Bi-LSTM** (Graves and Schmidhuber, 2005), **SMSD** (Xiong et al., 2019) which adopts self-matching networks

---

[2] `https://huggingface.co/bert-base-uncased`
[3] `https://github.com/lukemelas/PyTorch-Pretrained-ViT`

Table 2: Comparison results for sarcasm detection. † indicates ResNet backbone and ‡ indicates ViT backbone.

	Model	Acc(%)	P(%)	R(%)	F1(%)
Text	TextCNN	80.03	74.29	76.39	75.32
	Bi-LSTM	81.90	76.66	78.42	77.53
	SMSD	80.90	76.46	75.18	75.82
	BERT	83.85	78.72	82.27	80.22
Image	Image	64.76	54.41	70.80	61.53
	ViT	67.83	57.93	70.07	63.43
Multi-Modal	HFM†	83.44	76.57	84.15	80.18
	D&R Net†	84.02	77.97	83.42	80.60
	Att-BERT†	86.05	80.87	85.08	82.92
	InCrossMGs‡	86.10	81.38	84.36	82.84
	CMGCN‡	86.54	–	–	82.73
	Ours†	87.02	**82.97**	84.90	83.92
	Ours‡	**87.36**	81.84	**86.48**	**84.09**

and low-rank bilinear pooling for sarcasm detection, and **BERT** (Devlin et al., 2019) that generates predictions based on the [CLS] token as baseline models. For pure image-based models, we follow (Cai et al., 2019; Liang et al., 2021) to utilize the feature representations after the pooling layer of **ResNet** and [CLS] token in each image patch sequence obtained by **ViT** to generate predictions. For multi-modal based methods, we adopt **HFM** (Cai et al., 2019), **D&R Net** (Xu et al., 2020a), **Att-BERT** (Pan et al., 2020), **InCrossMGs** (Liang et al., 2021) and a variant of **CMGCN** (Liang et al., 2022) without external knowledge as the multi-modal baselines.

4.4 Results without External Knowledge

We first evaluate the effectiveness of our proposed framework by comparing with the baseline models as shown in Tabel 2. It is shown that our proposed model achieves state-of-the-art performance. Obviously, text-based models perform far better than image-based methods, which implies that text is more comprehensible and more informative than images. This supports our intuition of extracting textual knowledge from images as additional clues. On the other hand, multi-modal methods outperform all those models in single modality. This illustrates that considering information from both modalities contributes to the task by providing additional cues on modality associations.

Note that compared with multimodal methods using ResNet as the visual backbone network, our model achieves a 0.97% improvement in terms of Accuracy and a 1.00% improvement in terms of F1-score over the state-of-art method **Att-BERT**. Besides, using ViT as the image feature extractor, our model outperforms the **InCrossMGs** model with a 1.26% improvement in Accuracy and 1.25%

Table 3: Results of different knowledge types.

Knowledge Type	Acc(%)	F1(%)
w/o external knowledge	87.36	84.09
Image Attributes	86.43	83.30
ANPs	86.35	83.54
Image Captions	**88.26**	**84.84**
Image Captions (w/o image)	86.60	83.28

Table 4: Experimental results of ablation study.

Model	Acc(%)	F1(%)
Ours	**87.36**	**84.09**
w/o atomic-level	86.56	83.50
w/o MCA and atomic-level	86.01	82.73
w/o composition-level	82.60	79.13

improvement in F1-score. Our method can also achieve better performance with improvement of 0.82% based on Accuracy compared with the recent proposed **CMGCN**. The results demonstrate the effectiveness and superiority of our framework for sarcasm detection by modeling both atomic-level and composition-level cross-modality congruities in a hierarchical manner.

4.5 Results with External Knowledge

We then evaluate the effectiveness of our method by considering external knowledge. Table 3 reports the accuracy and F1-score for our proposed sarcasm detection method enhanced by considering different types of knowledge. By incorporating image captions, the performance further improves compared with the original model (w/o external knowledge). On the contrary, Image Attributes and ANPs bring negative effects and deteriorate the performance. We conjecture two possible reasons, 1) image attributes and ANPs can sometimes be meaningless or even noisy for identifying sarcasm, 2) image attributes and ANPs are rather short, lacking rich compositional information for our hierarchical model. Last but not the least, it is worth mentioning that only exploiting texts and captions in textual modality (Image Captions w/o image) without the visual modality also achieves superior performance compared with all multi-modal baselines in Table 2. Such observation illustrates that the pre-trained models such as CLIP and GPT-2 can provide meaningful external information for sarcasm detection.

4.6 Ablation Study

Impact of Different Components. We conduct ablation studies using ViT as the visual backbone network without external knowledge to further understand the impact of different components of our proposed method. To be specific, we consider three different scenarios, 1) remove the atomic-level con-
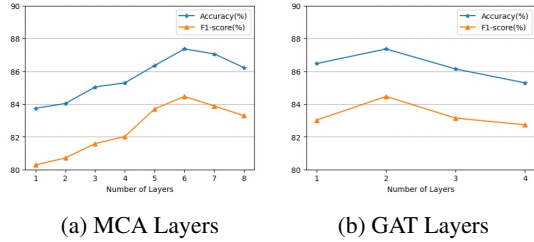
(a) MCA Layers   (b) GAT Layers

Figure 3: Performance of different model architectures.

gruity score $\mathbf{s}_a$ (denoted as w/o atomic-level), 2) remove both $\mathbf{s}_a$ and the multi-head cross attention (MCA) module by replacing $\tilde{T}$ to T in all the computations (denoted as w/o MCA and atomic-level), 3) remove composition-level congruity score $\mathbf{s}_p$ (denoted as w/o composition-level).

The results are shown in Table 4. It is clear that our model achieves the best performance when composing all these components. It is worth noting that the removal of composition-level $\mathbf{s}_p$ leads to significant performance drop, compared to atomic-level removal. This indicates that the composition-level congruity plays a vital role for discovering inconsistencies between visual and textual modalities by exploiting complex structures through propagating atomic representations along the semantic or geographical dependencies. Moreover, the removal of MCA leads to slightly lower performance which indicates that cross attention is beneficial for modeling cross modality interactions and reducing the modality gap in the representation space.

Impact of MCA Layers. We measure the performance change without external knowledge when varying the number of MCA layers from 1 to 8 in Figure 3a. As can be seen, the performance first increases along with the increasing number of layers and then decreases after 6 layers. This shows excessive MCA layers in atomic-level congruity module may overfit to textual and visual modality alignment instead of sarcasm detection.

Impact of GAT Layers. We analyse the impact of the number of GAT layers for our proposed model and report Accuracies and F1 scores in Figure 3b. The results show that the best performance can be achieved when using a two-layer GAT model and the performance further drops when increasing the number of layers. We conjecture the reason to be the over-smoothing issue of Graph Neural Networks when increasing the number of propagation layers, making different nodes indistinguishable.

Impact of Different Sentence Embeddings. We

Table 5: Experimental results of different sentence embedding.

Model	USE	Word Averaging	CLIP	Bert
Accuracy(%)	86.98	87.02	**88.10**	87.10

(a) Sarcasm   (b) Non-Sarcasm

Figure 4: Wrongly detected samples by our framework without external knowledge.

perform experiments using Universal Sentence Encoder (USE) (Cer et al., 2018), CLIP (Radford et al., 2021), CLS Token of Bert (Devlin et al., 2019), and Word Averaging to extract sentence embedding (i.e., $\mathbf{c}$ in Eq 6) as shown in Tabel 5. Although CLIP outperforms other methods by a slight margin, we prefer Word Averaging to keep our model concise and reduce the number of parameters.

4.7 Case Study

To further justify the effectiveness of external knowledge for sarcasm detection task, we provide case studies on the samples that are incorrectly predicted by our proposed framework with only text-image pair but can be accurately classified with the help of image captions. For example, intravenous injection depicted in Figure 4a can indicate a flu or hospitalization via image modeling, which aligns with *flu* or *hospital* in the text input. However, by generating an image caption expressing a bad mood indicated by *suffered*, it becomes easy to detect the sarcastic nature of this sample by contrasting *fun* in the text description and *suffer* in the image caption. As another example shown in Figure 4b, the image encoder only detects a human holding a torch without any contexts and wrongly predicts the sample as sarcasm because of the disalignment between the image and text description. By generating the image caption expressing an *olympic athlete*, the knowledge-fused model is able to detect the alignment and correctly classifies this sample. This reflects that by further utilizing CLIP (Radford et al., 2021) and GPT-2 (Radford et al., 2019) models pre-trained using large-scale data as an external knowledge source, the generated image captions are more expressive to understand some
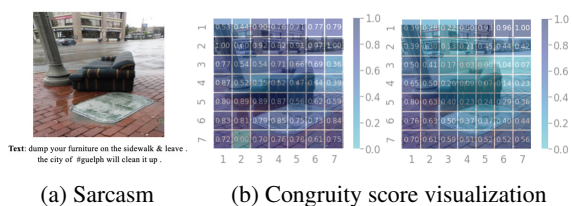
(a) Sarcasm   (b) Congruity score visualization

Figure 5: Visualization of atomic-level and composition-level congruity score between text and image. Better to be viewed in color format and zoom.

sophisticated visual concepts and to mitigate the furtiveness and subtlety of sarcasm.

We further illustrate the effectiveness of our hierarchical modeling by showing the congruity score maps in Figure 5. Given a sarcastic sample in Figure 5a, we visualize the congruity scores between the text and image in both atomic-level module $\mathbf{s}_a$ (left side of Figure 5b) and composition-level module $\mathbf{s}_p$ (right side of Figure 5b). The smaller the values, the less similar between the text and image (i.e., more likely to be detected as sarcasm). It can be shown that the atomic-level module attends to *furniture* in the image whereas the composition-level module down-weighs those patches, making the text and image less similar for sarcasm prediction. Correspondingly, our proposed hierarchical structure has the power to refine atomic congruity to identify more complex mismatches for multi-modal sarcasm detection using graph neural networks.

5 Conclusion

In this paper, we propose to tackle the problem of sarcasm detection by reasoning atomic-level congruity and composition-level congruity in a hierarchical manner. Specifically, we propose to model the atomic-level congruity based on the multi-head cross attention mechanism and the composition-level congruity based on graph attention networks. In addition, we propose to exploit the effect of various knowledge resources on enhancing the discriminative power of the model. Evaluation results demonstrate the superiority of our proposed model and the benefit of image captions as external knowledge for sarcasm detection.

Limitations

We present two possible limitations: 1) we only use the Twitter dataset for evaluation. However, to the best of our knowledge, this dataset is the only benchmark for the evaluation of multi-modal sar-

casm detection in our community. Nevertheless, we conduct extensive experiments with various metrics to show the superiority of our proposed method. We leave the construction of more high-quality benchmarks in our future work; 2) our knowledge enhancement strategy in Section 3.5 may not be suitable for ANPs and Image Attributes. We analyze the results in Section 4.5. Consequently, there is a pressing need for a more general and elegant knowledge integration method in view of the importance of external knowledge for multi-modality sarcasm detection.

Ethics Statement

This paper is informed by the ACM Code of Ethics and Professional Conduct. Firstly, we respect valuable and creative works in sarcasm detection and other related research domains. We especially cite relevant papers and sources of pre-trained models and toolkits exploited by this work as detailed and reasonable as possible. Besides, we will release our code based on the licenses of any used artifacts. Secondly, our adopted dataset does not include sensitive privacy individual information and will not introduce any information disorder to society. For precautions to prevent re-identification of data, we mask facial information in Figure 4b. At last, as our proposed sarcasm detection method benefits the identification of authentic intentions in multi-modal posts on social media, we expect our proposed method can also bring positive impact on related problems, such as opinion mining, recommendation system, and information forensics in the future.

ACKNOWLEDGEMENT

References

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2,*

2019, Volume 1: Long Papers, pages 2506–2515. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. pages 169–174. Association for Computational Linguistics.

D Davidov, O Tsur, and A Rappoport. 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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Aniruddha Ghosh and Tony Veale. 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*, pages 482–491, Copenhagen, Denmark. Association for Computational Linguistics.

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

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

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational*

Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers, pages 757–762. The Association for Computer Linguistics.

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

Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. 2021a. Sarcasm detection with commonsense knowledge. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3192–3201.

Ying Li, Hongwei Zhou, Yeyu Yin, and Jiaquan Gao. 2021b. Multi-label pattern image retrieval via attention mechanism driven graph convolutional network. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 300–308. ACM.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 4707–4715. ACM.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multimodal 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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1767–1777. Association for Computational Linguistics.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10918–10927. Computer Vision Foundation / IEEE.

Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. 2021. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 871–880. Association for Computational Linguistics.

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

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Linguistics*, 2:207–218.

Hao Sun, Yuntao Li, and Yan Zhang. 2022. Conlearn: Contextual-knowledge-aware concept prerequisite relation learning with graph neural network. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 118–126. SIAM.

Hao Sun, Zijian Wu, Yue Cui, Liwei Deng, Yan Zhao, and Kai Zheng. 2021. Personalized dynamic knowledge-aware recommendation with hybrid explanations. In *International Conference on Database Systems for Advanced Applications*, pages 148–164. Springer.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*, volume 215 of *Frontiers in Artificial Intelligence and Applications*, pages 765–770. IOS Press.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2115–2124. ACM.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 3777–3786. Association for Computational Linguistics.

Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. 2020b. Cross-modal attention with semantic consistence for image-text matching. *IEEE Trans. Neural Networks Learn. Syst.*, 31(12):5412–5425.

A Model Overview

For illustration, we give a figure of the text-image branch that can capture atomic-level and composition-level congruity between textual and visual modalities for multimodal sarcasm detection. Specifically, this figure is comprised of three main components: Feature Representation, Atomic-Level Cross Modality Congruity and Composition-Level Cross Modality Congruity, where Feature Representation extracts feature representations corresponding to texts and images, Atomic-Level Cross Modality Congruity obtains congruity scores via a multi-head cross-attention mechanism, and Composition-Level Cross Modality Congruity produces composition-level congruity scores based on constructed textual and visual graphs.

B Modal Parameters

For our model, the max length of sarcasm text is set to 100 and the max length of generated image caption is set to 20. For the architecture, the number of the multihead cross-attention layer is set to 6 for text-image branch and 3 for text-knowledge branch to capture atomic-level congruity score. The head
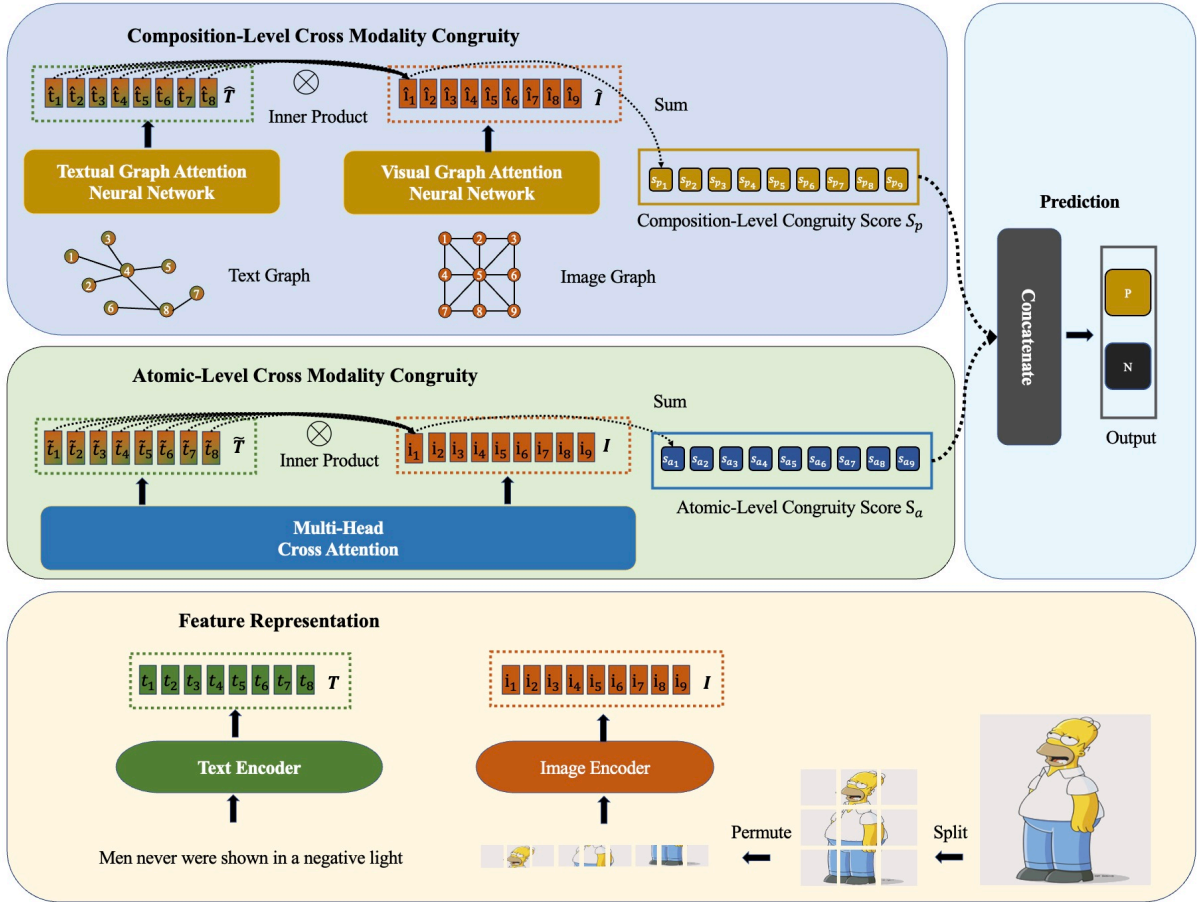
Figure 6: An overview of text-image branch of our approach, which involves three main modules: Feature Representation, Atomic-Level Cross Modality Congruity and Composition-Level Cross Modality Congruity.

Table 6: Parameters of our model

Parameters	Value
Max length of text	100
Max length of image caption	20
MCA layers for text-image branch	6
MCA layers for text-knowledge branch	3
Head number of MCA	5
GAT layers for text-image branch	2
GAT layers for text-knowledge branch	2
Batch size	32
Learning rate	$2e{-}5$
Weight decay	$5e{-}3$
Dropout rate	0.5

number is set to 5. The number of graph attention layer is set to 2 to obtain composition-level congruity score for both branches. We use Adam optimizer with a learning rate of $2e{-}5$, weight decay of $5e{-}3$, batch size as 32 and dropout rate as 0.5 to train the model. For more detail, please refer to the checklist in our implementation.