Information Fusion 100 (2023) 101921

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

KnowleNet: Knowledge fusion network for multimodal sarcasm detection

Tan Yue [a,b], Rui Mao [b], Heng Wang [a], Zonghai Hu [a], Erik Cambria [b,*]

[a] School of Electronic Engineering, Beijing Key Laboratory of Work Safety Intelligent Monitoring, Beijing University of Posts and Telecommunications, 10 Xitucheng Road, Haidian District, Beijing, 100876, China

[b] School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Ave, Block N4 #02a, Singapore, 639798, Singapore

ARTICLE INFO

Keywords:
Sarcasm detection
Multimodal learning
Information fusion

ABSTRACT

Sarcasm is a form of communication often used to express contempt or ridicule, where the speaker conveys a message opposite to their true meaning, typically intending to mock or belittle a specific target. Sarcasm detection has gained great attention in the field of natural language processing due to the fact that sarcasm is widespread on social media and difficult to detect for machines. While early efforts in sarcasm detection solely relied on textual data, the abundance of multimodal data on social media is also non-negligible. Recent research has focused on multimodal sarcasm detection, where attention mechanisms and graph neural networks were commonly used to identify relevant information in both image and text data. However, these methods may overlook the importance of prior knowledge and cross-modal semantic contrast, which are crucial factors for human sarcasm detection. In this paper, we propose a novel model named KnowleNet that leverages the ConceptNet knowledge base to incorporate prior knowledge and determine image–text relatedness through sample-level and word-level cross-modal semantic similarity detection. Contrastive learning is also introduced to improve the spatial distribution of sarcastic (positive) and non-sarcastic (negative) samples. The proposed model achieves state-of-the-art performance on publicly available benchmark datasets.

1. Introduction

According to the Macmillan English Dictionary, sarcasm is defined as *the activity of saying or writing the opposite of what you mean, or of speaking in a way intended to make someone else feel stupid or show them that you are angry* [1,2]. Sarcasm detection is critical for sentiment analysis, because the intended meaning of a sarcastic expression likely has an opposite sentiment polarity, compared to its literal meaning [3,4].

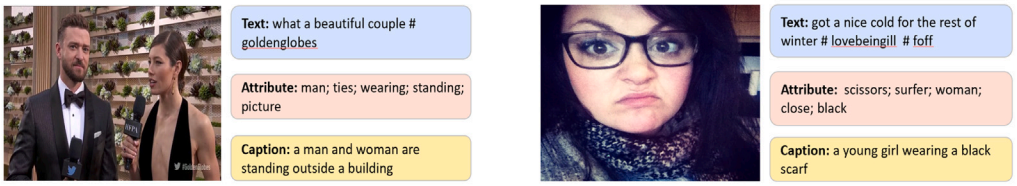
At an early stage, sarcasm detection mainly utilized textual information [5,6]. With the development of online social media, users upload a large amount of multimodal information to social media, e.g., image, audio, text, etc., which brings new challenges for sarcasm detection. Then, more and more researchers have drawn attention to multimodal sarcasm detection [7], whose models are based on attention mechanisms and graph neural networks. Some attention-based models capture the inconsistency between modalities by designing different deformation structures of attention mechanisms [8–10], while some graph neural network-based models build multimodal relationships with graphs by constructing cross-modal graph networks [11,12]. Some works [13–16] argued that external knowledge is critical for detecting and understanding non-literal expressions. The integration of commonsense knowledge has also proven instrumental in enhancing model performance in the field of affective computing [17,18], because the inherent difficulty in acquiring commonsense knowledge through task-specific dataset learning alone. However, the aforementioned sarcasm detection models boost performance by finding the emotional clues of images and text, and ignore the importance of commonsense knowledge for implicit emotion recognition.

In contrast, we intend to construct a model that fits the way of human sarcasm detection with knowledge. From the perspective of human sarcasm detection, prior knowledge and semantic similarity detection between modalities are very important. As shown in Fig. 1a, in non-sarcastic multimodal messages, the textual information is explicit, literally describing the content of the image. In other words, the visual and textual information of a non-sarcastic instance is strongly related. On the contrary, in the sarcastic example (Fig. 1b), semantic information in text and image is often contrastive or implicit. Thus, the image–text information is weakly related. The weak relatedness needs to be identified with some common-sense information, because implicit expressions of emotion often require more conceptual understanding of words in different situations. For example, "*nice cold*" is an expression that defies common sense, since being sick is a bad thing.

In this work, we propose a knowledge fusion network (KnowleNet). We first introduce the ConceptNet knowledge base [19] to obtain the conceptual knowledge, and then design a new multimodal information

* Corresponding author.
E-mail addresses: zhhu@bupt.edu.cn (Z. Hu), cambria@ntu.edu.sg (E. Cambria).

https://doi.org/10.1016/j.inffus.2023.101921
Received 4 April 2023; Received in revised form 6 July 2023; Accepted 7 July 2023
Available online 13 July 2023



(a) Non-sarcasm (b) Sarcasm

Fig. 1. Examples of non-sarcasm and sarcasm. Each sample includes text, image, image caption, and image attribute.

fusion method, i.e., cross-modal semantic similarity detection method, to detect the relatedness between image and text information. Due to the combination of contrastive learning methods, the spatial distribution of positive and negative samples is separated effectively [20], which leads to a significant improvement in classification performance.

We evaluate our model on the publicly available multimodal sarcasm detection benchmark dataset [21]. Compared with unimodal and multimodal strong baselines, our KnowleNet achieves the best accuracy (88.87%) and F1-score (86.33%), exceeding the strongest baseline by 1.64% in accuracy and 2.88% in F1. The above improvements demonstrate the effectiveness of our method. In addition, analytical experiments explain the effect of different modules in the KnowleNet. Our ablation study shows considerable performance degradation when we remove commonsense knowledge based semantic similarity detection module (−2.98%) and contrastive learning triplet loss module (−2.41%), which also proves the success of our proposed technical components.

The main contributions of this work are summarized as follows:

- We propose a novel knowledge fusion network for multimodal sarcasm detection, leveraging ConceptNet knowledge. To the best of our knowledge, it is the first model that combines prior knowledge to improve sarcasm detection accuracy.
- A new multimodal information fusion method is presented. The word-level and sample-level cross-modal semantic similarity detection modules are designed to detect the semantic consistency of different modal information, which could learn useful features for sarcasm detection.
- We introduce contrastive learning to distinguish the distribution of sarcastic (positive) and non-sarcastic (negative) samples. Triplet loss is used to improve the representation of multimodal features, yielding state-of-the-art performance on two publicly available benchmark datasets.

2. Related works

2.1. Multimodal learning

Modality refers to the specific way in which people receive information. Since multimedia data is often the transfer medium of multiple forms of information, multimodal learning has gradually developed as the main approach of multimedia content analysis and understanding [22]. Researchers have achieved remarkable research results in the field of multimodal learning [23,24].

Also, many image–text information fusion models have emerged in recent years. The ERNIE-VIL [25] model proposed by Baidu used the structured knowledge in the scene graph to enable the model to perform fine-grained semantic alignment. The VIVO [26] model used Image-Tag for pre-training so that the semantic tag could be aligned with the region features in the image, enabling it to be used in the Image Caption task to solve the problem of novel object recognition. RpBERT [27] used a multimodal BERT model for multimodal named entity recognition tasks, and the proposed relation propagation mechanism could make better use of visual information based on the relatedness between image and texts.

2.2. Multimodal sarcasm detection

In initial studies, sarcasm detection models mainly used textual information [28–31]. The text data-based models performed sarcasm detection by improving the language model with contextual semantic features, word frequency, symbols, etc. With the emergence of more and more multimodal data in social media, the use of multimodal data for sarcasm detection has gradually attracted the attention of researchers.

Multimodal sarcasm detection focuses more on information fusion, representation, and information relatedness among multiple modalities, which sets it apart from text-based models. Schifanella et al. [32] proposed the first multimodal sarcasm detection model, which combined textual and visual modalities using two different computational frameworks. Cai et al. [21] created a multimodal sarcasm detection dataset from Twitter and presented a hierarchical fusion model that used image attribute information. Recent models have focused on attention mechanisms and graph neural networks (GNN). For instance, Pan et al. [9] proposed a BERT-based model with inter-modality attention to capture intra and inter-modality inconsistency. Wang et al. [10] designed a 2D-Intra-Attention mechanism to extract relationships between words and images. Tomás et al. [33] proposed a transformer-based architecture for the fusion of textual and visual information. In terms of GNN, an interactive graph convolutional network (GCN) structure was explored to learn inconsistent relationships in-modal and cross-modal graphs in a joint and interactive way to identify important clues in sarcasm detection [11]. And Liang et al. [12] also designed a cross-modal graph convolutional network to make sense of the inconsistent relations between modalities for multimodal sarcasm detection. Recently, Malik et al. [34] analyzed whether image information is necessary to understand the sarcastic intent of the text.

2.3. Challenge and motivation

However, existing image–text fusion methods only focus on finding clues from the related information between images and text, and these methods often have some errors and poor generalization in finding implicit sarcasm clues. In addition, due to the lack of prior knowledge, these models do not have an understanding of commonsense information.

In order to re-think important features for sarcasm detection based on how humans recognize sarcasm, we adopt a novel perspective that emphasizes the significance of prior knowledge and cross-modal semantic similarity. Given that sarcastic content often involves a reversal of the expresser's actual intention, and that the same words or sentences may convey different meanings in sarcastic situations, the incorporation of prior knowledge with additional concepts is essential. In addition, modeling the relatedness of information across modalities through cross-modal semantic similarity detection is effective for detecting sarcasm. In contrast to existing methods, we introduce both word-level and sample-level cross-modal semantic similarity detection, and employ a contrastive learning loss function to express the similarity value as the distance between samples using a high-dimensional spatial distance metric.

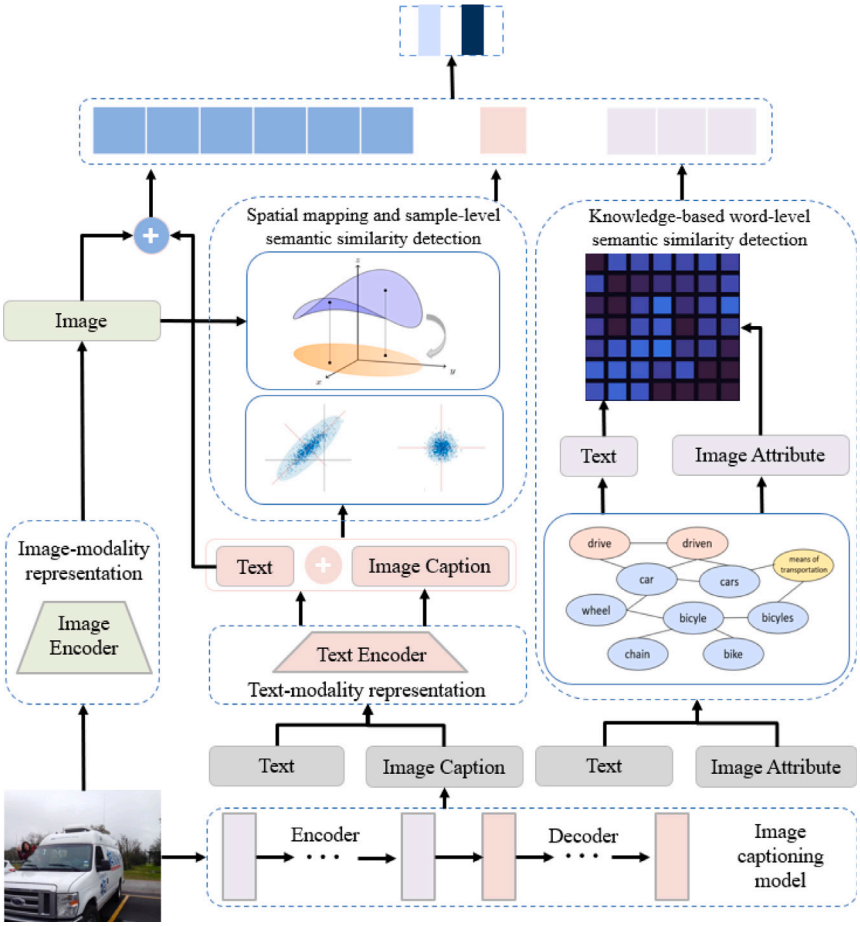


Fig. 2. Overview of our proposed knowledge fusion network.

3. Methodology

We present KnowleNet, a knowledge fusion network for multimodal sarcasm detection. The architecture of our proposed model is depicted in Fig. 2. We consider text, image, image attribute, and image caption as four modalities of data. Image attributes refer to the five attribute words extracted from the images, while image captions denote sentences generated from the images. Initially, we utilize a pre-trained image caption model to generate the captions and encode all the data to obtain feature representations. Specifically, we apply the BERT model to encode text and image captions, while the ResNet model is used to encode image data. Subsequently, we devise a knowledge-based word-level semantic similarity detection module and a sample-level semantic similarity detection module to extract effective features based on the semantic consistency of information across modalities. Furthermore, we introduce the contrastive learning loss function to improve the separation of the spatial distribution of positive and negative samples.

3.1. Text and image feature representation

We use the method proposed by Xu et al. [35] as the image caption model to generate captions. For the lower computation cost, we replace the encoder with pre-trained MobileNetV3 [36]. The images $\mathcal{I}$ are input to the image caption model to get the caption: $S_c = \{s_c^1, \ldots, s_c^l\}$ (l is the length of the caption).

For the text data, we express it as $S_t = \{s_t^1, \ldots, s_t^m\}$ (m is the length of the sequence). The BERT model embeds and encodes S_c and S_t to yield their representations. The BERT encoder's outputs are the *pooled_output* to represents each input sequence as a whole. ($T_b, C \in \mathbb{R}^{1\times d_b}$; d_b is the dimension of BERT hidden states.)

$$T_b = \mathcal{M}^{BERT}(S_t), \tag{1}$$

$$C = \mathcal{M}^{BERT}(S_c), \tag{2}$$

where $\mathcal{M}^{BERT}$ is the model of BERT.

As the same as the text, the images $\mathcal{I}$ are put into the pre-trained ResNets model for feature extraction and further processed by the average pooling layer. The image representation ($I \in \mathbb{R}^{1\times d_r}$; d_r is the dimension of ResNet hidden states) is given by

$$I = AverP(\mathcal{M}^{ResNet}(\mathcal{I})), \tag{3}$$

where $\mathcal{M}^{ResNet}$ is the model of ResNet, and $AverP$ is the average pooling operator.

3.2. Knowledge-based word-level semantic similarity detection

Semantic consistency between image and text is often a key feature in determining sarcasm. As shown in Fig. 3, unlike the existing models based on attention mechanism, we first introduce the ConceptNet knowledge network to associate more similar concepts, i.e., the prior knowledge, with image attributes and text data.

ConceptNet [19] is a semantic network that aids computers in comprehending the meaning of words used by people. This network is represented by a sparse, symmetrical matrix. Each word is represented by other words connected in the ConceptNet. The network calculates the pointwise mutual information of the matrix entries, which is smoothed with contextual distribution. Negative values are clipped to

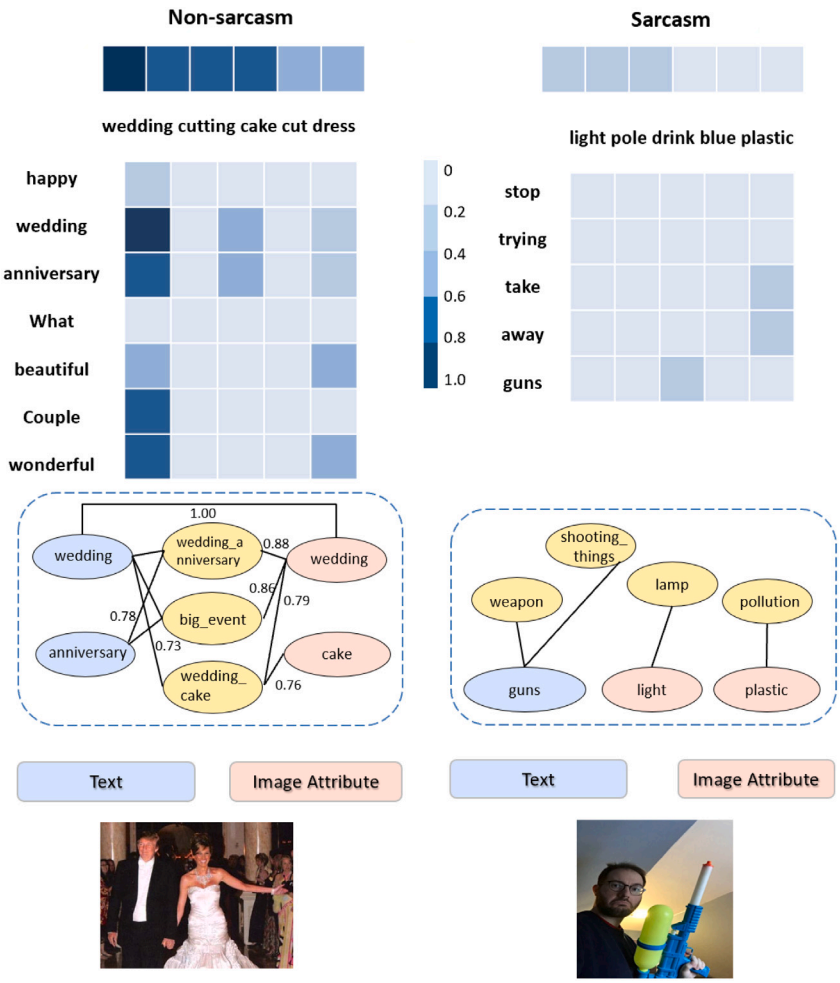


Fig. 3. Different processing procedures for positive(sarcasm) and negative(non-sarcasm) samples in knowledge-based word-level cross-modal semantic similarity detection module.

generate positive pointwise mutual information (PPMI). The dimension of the resulting matrix is reduced to 300 dimensions through truncated SVD, and the terms and contexts are combined symmetrically into a single matrix of word embeddings. After that, the new vectors $\hat{q}_i$ are updated by minimizing the objective function $\Psi(Q)$ to be close to their original values q_i, and to the adjacent words in the graph with edge E.

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \left\| q_i - \hat{q}_i \right\|^2 + \sum_{i,j \in E} \beta_{ij} \left\| q_i - q_j \right\|^2 \right], \tag{4}$$

where, α_i and β_{ij} are the connection weight values in ConceptNet. When the original vector values do not exist, α_i is set to 0. The above yields the word embedding matrix of ConceptNet with a lexicon size of 510K.

We express the image attribute as $S_a = \{s_a^1, \ldots, s_a^u\}$ (u is the length of the sequence). We use the ConceptNet to process S_a and S_t to get their vectorized representations ($T_c \in \mathbb{R}^{k \times d_h}, A \in \mathbb{R}^{u \times d_h}$; d_h is the dimension of ConceptNet.):

$$T_c = ConceptNet(S_t), \tag{5}$$

$$A = ConceptNet(S_a). \tag{6}$$

The text features and image attributes for which conceptual knowledge is obtained are used to determine similarity by matrix calculation of the inner product. Since the sparsity of the matrix introduces additional interference, we use a Max Pooling operator to extract key features, followed by a Flatten layer for the representation.

$$F_{ta} = Flat(MaxP\left(T_c \otimes A^T\right)), \tag{7}$$

where $MaxP$ denotes the Max Pooling operator, $Flat$ means the Flatten layer.

Since the image attributes are extracted from images, which represent the information in the image to a great extent. We calculate the similarity between image and text information by the method of word-level semantic similarity detection, and more image–text association clues can be found due to the introduction of concept knowledge.

3.3. Spatial mapping and sample-level semantic similarity detection

In addition to word-level semantic similarity detection, we consider the sample level. We consider that not only the degree of relatedness between words in the text and words in the image attributes but also the feature information of the whole image has some relatedness with the whole text, which is also important for the calculation of semantic consistency. However, text and images have different encoders, resulting in their feature vectors not being in the same coordinate system. If the image and text feature vectors are simply computed, satisfactory results cannot be obtained.

Observing this problem, we perform matrix coordinate transformations on the feature vectors of text and images. Specifically, we first concatenate the feature vectors of the image and the text data. And then, we centralize the feature vectors, i.e., subtract the mean values. ($X \in \mathbb{R}^{(d_r + d_b * 2) \times n}$; n is the number of samples)

$$T_{tc} = T_b \oplus C, \tag{8}$$

$$I = \left\{m_1, m_2 \ldots . m_n\right\}, \bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i, \tag{9}$$

$$T_{tc} = \{t_1, t_2 .... t_n\}, \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i. \tag{10}$$

$$X = (T_{tc} \oplus I)^T = \begin{pmatrix} t_1, & t_2, & \cdots & t_n \\ m_1, & m_2, & \cdots & m_n \end{pmatrix}, \tag{11}$$

Then the covariance matrix with its eigenvalues and the corresponding eigenvectors are derived.

$$\begin{aligned} A &= \frac{1}{n-1} X X^T \\ &= \begin{pmatrix} \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2 & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})(m_i - \bar{m}) \\ \vdots & \ddots & \vdots \\ \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})(m_i - \bar{m}) & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}(m_i - \bar{m})^2 \end{pmatrix}. \end{aligned} \tag{12}$$

The covariance matrix A of X can be calculated above, and then the eigenvalues and eigenvectors are solved by the eigenvalue decomposition method to obtain Q and $\sum$. Q is a matrix composed of eigenvectors of matrix A, $\sum$ is a diagonal array, and the elements on the diagonal are the eigenvalues. We take the first p columns of Q as the transformation matrix $P \in \mathbb{R}^{(d_r+d_b*2)\times p}$, then $Y_t = T_{tc}W_1P$; $Y_i = IW_2P$. ($Y_t, Y_i \in \mathbb{R}^{1\times p}$.) W is used to change the feature vectors of text and images into the same dimension as P. ($W_1 \in \mathbb{R}^{d_b*2\times(d_r+d_b*2)}$, $W_2 \in \mathbb{R}^{d_r\times(d_r+d_b*2)}$.)

$$A = Q \sum Q^{-1} \tag{13}$$

$$D = (Y_t - Y_i)(Y_t - Y_i)^T. \tag{14}$$

With the feature matrix P, we can obtain the vectors Y_t and Y_i that have been dimensionally reduced and mapped to the same coordinate space. Then, we calculate the spatial distance of the feature vectors D to represent the semantic similarity of the image and text information.

Finally, we use the dropout layer to avoid overfitting problems and use multiple fully connected layers to further fuse each modal feature and classification.

$$T_{f1} = Drop(T_b), \tag{15}$$

$$T_{f2} = Drop(\sigma(FC(I \oplus C))), \tag{16}$$

$$\hat{\mathbf{y}} = Sigmoid(\sigma(FC(T_{f1} \oplus T_{f2}) \oplus \sigma(FC(F_{ta})) \oplus \sigma(FC(D)), \tag{17}$$

where $Drop$ denotes the Dropout layer, and FC denotes the fully connected layer. σ is the $ReLU$ activation function. $\oplus$ denotes the concatenate operator.

3.4. Loss function

Furthermore, we introduce a contrastive learning loss function to conduct the spatial distribution differentiation of positive and negative samples. BinaryCrossentropy loss function and Triplet loss function are combined for learning. The BinaryCrossentropy loss learns to classify sarcastic and non-sarcastic samples in vector space. The Triplet loss function learns to expand the distance between positive and negative examples in vector space [37].

Specifically, given a triplet (x, x^+, x^-) of a data set with sample size n. As shown in Eq. (18), we aim to minimize the distance between anchor points and positive samples while maintaining the distance between them and negative samples. So that features with the same label are as close as possible in spatial location, while features with different labels are as far away as possible in spatial location. At the same time, in order to keep the features of the samples from aggregating into a very small space, it is required that the negative example should be more distant than the positive example from the anchor point.

$$d\left(x, x^+\right) \to 0,\ d\left(x, x^-\right) \to d\left(x, x^+\right) + \alpha\ , \tag{18}$$

where the $d(\cdot,\cdot)$ denotes the distance between two points and α denotes the margin.

Table 1
Detailed information of the multimodal sarcasm detection benchmark datasets.

	Training	Development	Test
Positive (Sarcasm)	8642	959	959
negative (Non-Sarcasm)	11 174	1451	1450
Total	19816	2410	2409
Token Length	16.91	16.92	17.13

The overall distance of the triplet is expressed as

$$L = \max\left\{d\left(x, x^+\right) - d\left(x, x^-\right) + \alpha, 0\right\} \tag{19}$$

According to the above equation, our input as $h\left(x_i\right)$, where $h\left(x_i\right)$ represents semantic similarity features of word-level and sample-level, as shown in Fig. 4 and Eq. (20).

$$h\left(x_i\right) = \sigma(w_t F_{ta} + b_t) \oplus \sigma(w_d D + b_d), \tag{20}$$

$$\forall\left(h\left(x_i^a\right), h\left(x_i^p\right), h\left(x_i^n\right)\right) \epsilon\ \gamma, \tag{21}$$

where the w_t, w_d denote the linear weight and b_t, b_d denote the bias. γ is a triplet with sample size n.

Therefore, the specific calculation can be expressed as

$$\left\|h\left(x_i^a\right) - h\left(x_i^p\right)\right\|_2^2 + \alpha < \left\|h\left(x_i^a\right) - h\left(x_i^n\right)\right\|_2^2, \tag{22}$$

The loss function is minimized as

$$L_c = \sum_{i}^{n}\left[\left\|h\left(x_i^a\right) - h\left(x_i^p\right)\right\|_2^2 - \left\|h\left(x_i^a\right) - h\left(x_i^n\right)\right\|_2^2 + \alpha\right]. \tag{23}$$

The corresponding gradient calculation is

$$\begin{aligned} \frac{\partial L_c}{\partial h(x_i^a)} &= 2(h(x_i^n) - h(x_i^p)), \\ \frac{\partial L_c}{\partial h(x_i^p)} &= 2(h(x_i^p) - h(x_i^a)), \\ \frac{\partial L_c}{\partial h(x_i^n)} &= 2(h(x_i^a) - h(x_i^n)). \end{aligned}$$

The total loss function can be expressed as (BCE denotes the BinaryCrossentropy loss function.)

$$L = L_b + L_c = \mathbf{BCE}\left(\hat{y}, y\right) + L_c. \tag{24}$$

4. Experiment

4.1. Dataset

We use the publicly available multimodal sarcasm detection benchmark dataset called Dataset-1 in our experiments. The dataset was created by Cai et al. [21] and contained English tweets. In addition, Cai et al. [21] also processed the dataset to extract image attributes from images. Therefore, each sample includes the image, text, and image attributes. The detailed data of the dataset is shown in Table 1.

Another multimodal sarcasm detection dataset, called Dataset-2 in our experiments, was created by Maity et al. [38]. They collected 5854 samples from open-source Twitter and Reddit platforms and the dataset also consisted of two modalities, text and image.

4.2. Baselines

We compare our proposed KnowleNet with multiple well-known existing methods, such as:

- **Image-modality methods:** These models use only image data for sarcasm detection. **ResNet** [39] is CNN-based image classifier with residual connections. **ViT** is the model proposed by Dosovitskiy et al. [40] to apply transformer to image classification. **ConvNeXt** [41] based on CNN and attracted a great deal of attention because of its state-of-the-art performance in image processing.

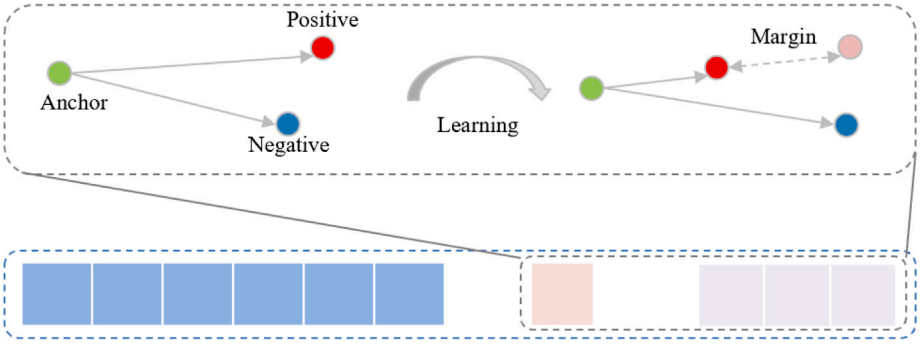


Fig. 4. The effectiveness of minimizing the triplet loss function.

- **Text-modality methods:** These models are based only on text data for sarcasm detection. **TextCNN** [42] based on convolutional neural network for text classification. The **Bi-LSTM** [43] structure is a type of recurrent neural network that can incorporate both past and future contextual information for predicting the current output. Meanwhile, **SIARM** [44] employs inner-attention mechanisms to detect sarcasm in textual data. **SMSD** [45] uses a self-matching network to capture semantic inconsistent information. **BERT** [46] is a pre-trained language model based on transformer.
- **Multimodal methods:** These models use both image and text data for multimodal sarcasm detection. **HFM** [21] proposed a hierarchical multimodal features fusion model for multimodal sarcasm detection. **VisualBERT** [47] is a pre-trained image–text model, consisting of a stack of transformer layers. Recent models like **D&R Net** [8], **Res-Bert** [9], **Intra-Att** [10], **Att-Bert** [9] based on attention mechanism, while **InCrossMGs** [11] and **CMGCN** [12] based on graph neural networks.

4.3. Settings

For baseline models, we use their default parameter settings as in the original papers or implementations. For the KnowleNet model, we respectively use a pre-trained ResNet-152 model and a BERT-base uncased model to encode image and text data. The dropout rate is 0.5. The length parameters are $k = 20, u = 5$. The dimension parameters are $d_b = 768, d_r = 2048, d_h = 300, p = 300$. We train the model with Adam [48] optimizer and stop training if the validation loss does not decrease for 10 consecutive epochs. The learning rate is $1e-5$.

4.4. Evaluation

Following existing baseline models, we use **Accuracy**, **Precision**, **Recall**, and **F1-score** to measure the model performance. Since the label distribution of the dataset is imbalanced, following Liang et al. [12], we also calculate the Macro-F1 score for evaluation. The Macro-F1 score is calculated by averaging the F1 values of each label, which means that the number of data is not considered, and each class is treated equally.

- **True Positives (TP):** the number of samples for which the prediction is positive, and the actual label is also positive.
- **False Positives (FP):** the number of samples for which the prediction is positive, and the actual label is negative.
- **True Negatives (TN):** the number of samples for which the prediction is negative, and the actual label is also negative.
- **False Negatives (FN):** the number of samples for which the prediction is negative, and the actual label is positive.

Table 2
Comparison of accuracy and F1-score results between the proposed KnowleNet model and other strong existing models on the Dataset-1. The models with $*$ are based on BERT and ResNet models.

Modality	Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
Text	TextCNN [42]	80.03	74.29	76.39	75.32
	SIARN [44]	80.57	75.55	75.70	75.63
	SMSD [45]	80.90	76.46	75.18	75.82
	Bi-LSTM [43]	81.90	76.66	78.42	77.53
	BERT [46]	83.85	78.72	82.27	80.22
Image	ResNet [39]	64.76	54.41	70.80	61.53
	ViT [40]	67.83	57.93	70.07	63.43
	ConvNeXt [41]	67.78	58.91	69.15	63.62
Multimodal	HFM [21]	83.44	76.57	84.15	80.18
	D&R Net [8]	84.02	77.97	83.42	80.60
	Res-Bert* [9]	84.80	77.80	84.15	80.85
	Intra-Att* [10]	85.64	80.01	**84.84**	82.35
	Att-Bert* [9]	86.05	78.63	83.31	80.90
	InCrossMGs [11]	86.10	81.38	84.36	82.84
	CMGCN* [12]	87.23	–	–	83.45
	KnowleNet*	**88.87**	**88.59**	84.18	**86.33**

$$Precision = \frac{TP}{TP + FP} \tag{25}$$

$$Recall = \frac{TP}{TP + FN} \tag{26}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{27}$$

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{28}$$

5. Results

In this section, we evaluate our KnowleNet with some strong baseline models to demonstrate the effect of our model.

Specifically, we want to show:

- Can our model achieve satisfactory results in the publicly available multimodal sarcasm detection benchmark dataset compared with unimodal methods and other multimodal methods?
- The KnowleNet model consists of many modules, what are the effects of each of these modules on KnowleNet?

5.1. Main results

On the Dataset-1, we divided the experiment into three groups. As shown in Table 2. **Text-modality methods:** BERT reaches the highest accuracy(83.85%) and F1-score(80.22%) based only on text data, which shows its outstanding performance in the sarcasm detection task. **Image-modality methods:** ConvNeXt and ViT achieve the accuracy of 67.78% and 67.83%, which are pretty close. However, compared with text-modality methods, the image-modality methods do not perform

Table 3
Comparison of Macro-F1 results between the proposed KnowleNet model and other baseline models on the Dataset-1. The models with ∗ are based on BERT and ResNet models.

Modality	Method	F1(%)	Macro-F1		
			Pre(%)	Rec(%)	F1(%)
Text	TextCNN [42]	75.32	78.03	78.28	78.15
	SIARN [44]	75.63	80.34	78.81	79.57
	SMSD [45]	75.82	80.87	78.20	79.51
	Bi-LSTM [43]	77.53	80.97	80.13	80.55
	BERT [46]	80.22	81.31	80.87	81.09
Image	ResNet [39]	61.53	60.12	73.08	65.97
	ViT [40]	63.43	65.68	71.35	68.40
	ConvNeXt [41]	63.62	64.85	72.73	68.56
Multimodal	HFM [21]	80.18	79.40	82.45	80.90
	Res-Bert* [9]	80.85	78.87	84.46	81.57
	Att-Bert* [9]	80.90	80.87	85.08	82.92
	InCrossMGs [11]	82.84	85.39	85.80	85.60
	CMGCN* [12]	83.45	–	–	85.61
	KnowleNet*	**86.33**	**88.83**	**88.21**	**88.51**

Table 4
Comparison of accuracy and F1-score results between the proposed KnowleNet model and other transformer-based models on the Dataset-1.

Model	Modality	Acc(%)	Pre(%)	Rec(%)	F1(%)
VisualBERT [47]	Multimodal	83.51	76.66	82.94	79.68
BERT [46]	Text	83.85	78.72	82.27	80.22
ViLBERT [49]	Multimodal	84.68	77.52	86.37	81.71
RoBERTa [50]	Text	88.28	86.32	85.48	85.89
BERTweet [51]	Text	88.36	87.26	88.01	87.63
ALBERT [52]	Text	89.17	88.86	87.65	88.25
KnowleNet (ALBERT-based)	Multimodal	92.69	91.57	90.85	91.21

well, which shows text data may contain more effective feature information. It will be more challenging if we rely only on image data for sarcasm detection. **Multimodal methods:** we compare our KnowleNet model with some strong models proposed for sarcasm detection or other multimodal tasks. Compared with unimodal methods, most multimodal methods perform better. However, with the continuous improvement of pre-trained models for text and image processing, some unimodal methods are now surpassing the performance of previous multimodal methods, as seen in the difference in performance of BERT(83.85%) and HFM(83.44%). Therefore, we use the same text and image pre-trained models for a fair comparison. As shown in Table 2, the models with ∗ are all based on BERT and ResNet models. The experimental results show that our model achieves the best results, outperforming the strongest baseline (CMGCN) by 1.64% in accuracy and 2.88% in F1, which demonstrates the effectiveness of our model.

In addition, because of the problem of positive and negative sample imbalance, we also introduce Macro-F1. As shown in Table 3, macro metrics are higher, which indicates that the problem of unbalanced positive and negative samples can obviously affect the final results, and the model has better detection performance for negative samples. Our KnowleNet also achieves better Macro-F1 results of 88.51%, which exceeds the strongest CMGCN model by 2.9%.

Recently, many transformer-based models have shown promising results. Therefore, we test multiple unimodal and multimodal transformer-based models on the Dataset-1. As shown in the Table 4, ViLBERT extends the BERT architecture to a multimodal model and get the accuracy of 84.68%. The accuracy of RoBERTa, BERTweet, and ALBERT is very close, reaching 88.28%, 88.36%, and 89.17%, respectively. The greater performance of ALBERT also helps our model to get higher accuracy (92.69%) and F1-score (91.21%).

Table 5
Comparison of accuracy and F1-score results on the Dataset-2.

Modality	Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
Text	BERT-GRU [38]	59.72	–	–	59.12
	RoBERTa [50]	61.82	62.03	60.31	61.16
Image	ResNet [39]	59.39	–	–	57.79
Multimodal	Maity et al. [38]	62.20	–	–	61.47
	KnowleNet	**64.35**	**63.72**	**62.08**	**62.89**

Most of existing works have conducted extensive experiments and comparisons on the Dataset-1. We also test our proposed model on an additional publicly available dataset for multimodal sarcasm detection. As shown in the Table 5, our KnowleNet achieves the best accuracy (64.35%) and F1-score (62.89%). Compare with the strongest baseline model proposed by Maity et al. [38], which also use the BERT model and the ResNet model as the encoders, our accuracy is improved by 2.15%.

5.2. Visualization study

For a better explanation of the effectiveness of our proposed model, we performed the visualization of the KnowleNet. As shown in Fig. 3, Fig. 4, and Fig. 5, the processing of the samples after they enter the model is shown in detail. In Fig. 5, it can be clearly observed that after the optimization of the contrastive learning loss function, the positive and negative samples are well separated in the spatial distribution. We use the technique of T-SNE to downscale the high-dimensional features (Eq. (20)) into three dimensions for visualization.

5.3. Ablation study

The proposed KnowleNet consists of many modules, since the combination of different modules in a model can produce different performances, we conduct some ablation experiments on Dataset-1 to show the effects of different modules.

As described in Section 3, the knowledge-based word-level semantic similarity detection module, the spatial mapping and sample-level semantic similarity detection module, and the triplet loss module are important modules in our model. As shown in Table 6, we remove different modules respectively and conduct ablation experiments. The performance degradation is considerable when we remove the spatial mapping and sample-level semantic similarity detection module (w/o S). This demonstrates the effectiveness of sample-level semantic similarity detection in adding useful feature information, as well as the usefulness of the matrix coordinate transformation technique in better fusing multimodal feature vectors generated by different encoders. Note that removing the knowledge-based word-level semantic similarity detection module (w/o $\mathcal{W}$) dramatically degrades the performance, which verifies the significance of the conceptual knowledge and our knowledge fusion method. From the results of w/o $\mathcal{L}$, we consider that the triplet loss is important to separate positive and negative samples and then affects the final result. This conclusion is also well demonstrated in Section 5.2.

5.4. Case study

To further explain the effectiveness of our KnowleNet for sarcasm detection task, we provide case study on the sample that is incorrectly predicted by Cai et al. [21] and is presented in their error analysis section. As shown in Fig. 6, the insulting gesture in Fig. 6b is in contrast to the text content ‘thanks for’. But the attention-based model could not obtain the commonsense information that this gesture is

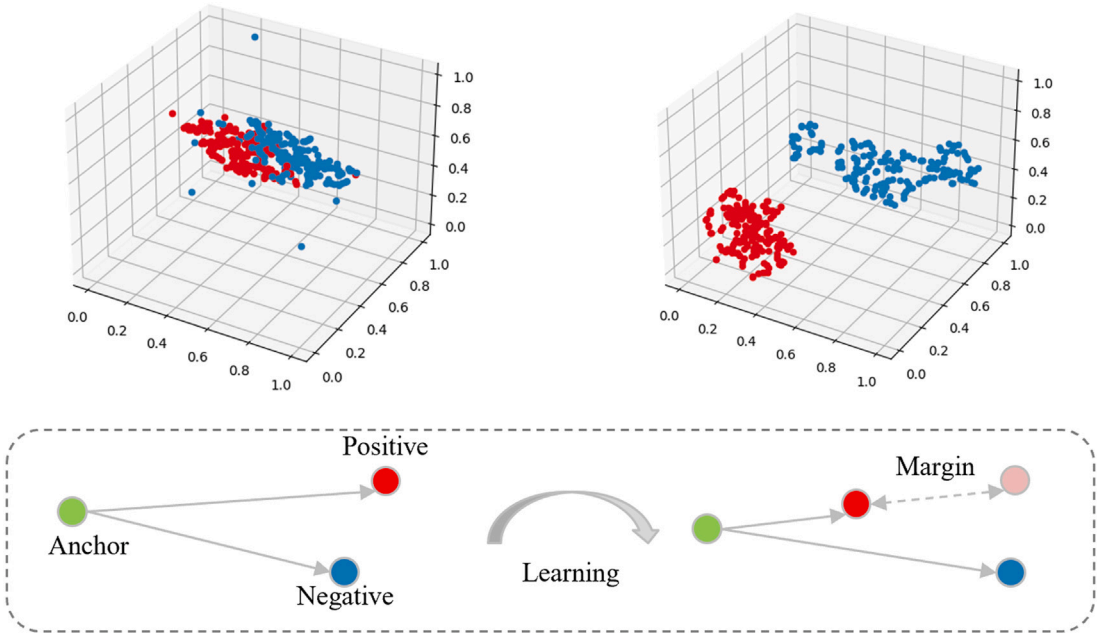


Fig. 5. Visualization results for the triplet loss optimization.

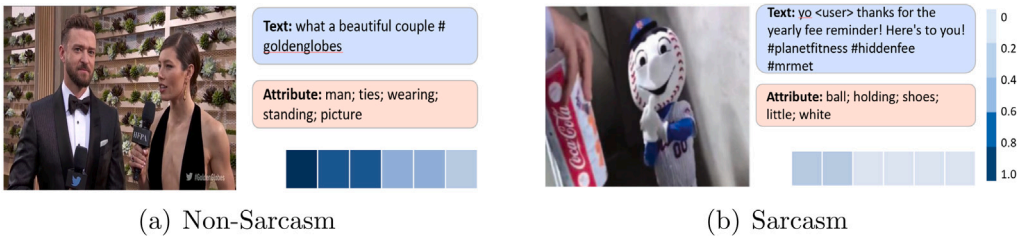


Fig. 6. Case study presentation for samples processed by the cross-modal semantic similarity detection module.

Table 6
Results of ablation experiments on Dataset-1. "w/o $\mathcal{S}$" means KnowleNet model without the mapping and sample-level semantic similarity detection module. "w/o $\mathcal{W}$" means KnowleNet model without the knowledge-based word-level semantic similarity detection module. "w/o $\mathcal{L}$" means KnowleNet model without the triplet loss module.

Model	Acc(%)	Δ	F1-score(%)	Δ	Macro-F1(%)	Δ
w/o $\mathcal{S}$	85.12	−3.75	81.85	−4.48	83.38	−5.13
w/o $\mathcal{W}$	85.89	−2.98	82.21	−4.12	84.58	−3.93
w/o $\mathcal{L}$	86.46	−2.41	83.32	−3.01	85.63	−2.88
KnowleNet	88.87	–	86.33	–	88.51	–

insulting, which causes the detection to fail. Our model obtains effective features by combining commonsense and conceptual knowledge and using cross-modal semantic similarity detection methods. We can observe that with the semantic similarity detection module, different features are generated for sarcastic and non-sarcastic samples, which will significantly impact the performance of sarcasm detection.

5.5. Summary of experimental results

In this section, we summarize the experimental results to demonstrate the validity of the proposed model. **(1).** By comparing the results of the unimodal and multimodal models, we observed that the multimodal model tends to perform better. Even though the results of the models based on image data are unsatisfactory, multimodal models with additional image data tend to outperform the models based on text data, benefiting from effective multimodal fusion methods. **(2).** Our model achieves better results when comparing existing multimodal and unimodal models. It is demonstrated that introducing conceptual knowledge (prior knowledge) followed by semantic similarity detection at different levels is effective for sarcasm detection. **(3).** Visualization and ablation experiments explain the effect of different modules in the KnowleNet. The word-level and sample-level semantic similarity detection methods could learn more effective features, while the triplet loss further optimizes the spatial distribution of positive and negative sample features.

6. Conclusion

A new and innovative model called Knowledge Fusion Network (KnowleNet) is proposed in this paper. We analyze the sarcasm detection task from a new perspective, i.e., by combining prior knowledge and the semantic similarity between image and text for determining sarcasm. Our word-level and sample-level cross-modal semantic similarity detection methods leverage conceptual knowledge information, which is a novel approach that has not been explored before to the best of our knowledge. Moreover, we employ a contrastive learning approach with the triplet loss to optimize the spatial distribution of positive and negative sample features. Our proposed model achieves state-of-the-art results on publicly available benchmark datasets, demonstrating its superior performance in multimodal sarcasm detection.

On the other hand, we find that metaphors frequently appeared in sarcastic expressions, because they allow for a layer of indirectness and irony. By using metaphors, speakers can express their true intent in a subtle or veiled manner. This creates a sense of incongruity between

what is said and what is meant, adding depth and complexity to the sarcastic expression. In future work, we will combine the proposed KnowleNet model and a metaphor processing tool [53] to study the relationships between metaphors and sarcastic expressions.

CRediT authorship contribution statement

Tan Yue: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Rui Mao:** Conceptualization, Methodology, Writing – review & editing. **Heng Wang:** Methodology, Formal analysis, Writing – original draft. **Zonghai Hu:** Resources, Writing – review & editing, Supervision, Funding acquisition. **Erik Cambria:** Resources, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work described in this paper is supported by the BUPT innovation and entrepreneurship support program (2022-YC-S002) and the China Scholarship Council (CSC) under Grant 202206470036.

References

[1] D.E. Shaffer, Macmillan english dictionary for advanced learners, Korea TESOL J. 5 (1) (2002) 183–187.
[2] S. Dews, E. Winner, Muting the meaning a social function of irony, Metaphor Symb. 10 (1) (1995) 3–19.
[3] E. Cambria, S. Poria, A. Gelbukh, M. Thelwall, Sentiment analysis is a big suitcase, IEEE Intell. Syst. 32 (6) (2017) 74–80, http://dx.doi.org/10.1109/MIS.2017.4531228.
[4] S. Frenda, A.T. Cignarella, V. Basile, C. Bosco, V. Patti, P. Rosso, The unbearable hurtfulness of sarcasm, Expert Syst. Appl. 193 (2022) 116398.
[5] M. Bouazizi, T. Ohtsuki, Sarcasm detection in Twitter: "all your products are incredibly amazing!!!" - Are they really? in: 2015 IEEE Global Communications Conference, GLOBECOM, IEEE, 2015, pp. 1–6.
[6] T. Ptáček, I. Habernal, J. Hong, Sarcasm detection on Czech and English Twitter, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 213–223.
[7] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.
[8] N. Xu, Z. Zeng, W. Mao, Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3777–3786.
[9] H. Pan, Z. Lin, P. Fu, Y. Qi, W. Wang, Modeling intra and inter-modality incongruity for multi-modal sarcasm detection, in: Findings of the Association for Computational Linguistics, EMNLP 2020, 2020, pp. 1383–1392.
[10] X. Wang, X. Sun, T. Yang, H. Wang, Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data, in: Proceedings of the First International Workshop on Natural Language Processing beyond Text, 2020, pp. 19–29.
[11] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, R. Xu, Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4707–4715.
[12] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, R. Xu, Multi-modal sarcasm detection via cross-modal graph convolutional network, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 1767–1777.
[13] R. Mao, C. Lin, F. Guerin, Word embedding and WordNet based metaphor identification and interpretation, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1, ACL, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1222–1231, http://dx.doi.org/10.18653/v1/P18-1113.
[14] K. Zhang, Y. Li, J. Wang, E. Cambria, X. Li, Real-time video emotion recognition based on reinforcement learning and domain knowledge, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1034–1047.
[15] M. Ge, R. Mao, E. Cambria, Explainable metaphor identification inspired by conceptual metaphor theory, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, no. 10, 2022, pp. 10681–10689, http://dx.doi.org/10.1609/aaai.v36i10.21313.
[16] R. Mao, X. Li, M. Ge, E. Cambria, MetaPro: A computational metaphor processing model for text pre-processing, Inf. Fusion (ISSN: 1566-2535) 86–87 (2022) 30–43, http://dx.doi.org/10.1016/j.inffus.2022.06.002.
[17] E. Cambria, Q. Liu, S. Decherchi, F. Xing, K. Kwok, SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis, in: LREC, 2022, pp. 3829–3839.
[18] W. Li, L. Zhu, R. Mao, E. Cambria, SKIER: A symbolic knowledge integrated model for conversational emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, no. 11, 2023, pp. 13121–13129, http://dx.doi.org/10.1609/aaai.v37i11.26541.
[19] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31, no. 1, 2017.
[20] L. Zhu, W. Li, R. Mao, V. Pandelea, E. Cambria, PAED: Zero-shot persona attribute extraction in dialogues, in: Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL, 2023.
[21] Y. Cai, H. Cai, X. Wan, Multi-modal sarcasm detection in Twitter with hierarchical fusion model, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2506–2515.
[22] Y. Xia, L. Zhang, Z. Liu, L. Nie, X. Li, Weakly supervised multimodal kernel for categorizing aerial photographs, IEEE Trans. Image Process. 26 (8) (2017) 3748–3758, http://dx.doi.org/10.1109/TIP.2016.2639438.
[23] R. Cadene, H. Ben-younes, M. Cord, N. Thome, MUREL: Multimodal relational reasoning for visual question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019.
[24] J. Zhu, Y. Zhou, J. Zhang, H. Li, C. Zong, C. Li, Multimodal summarization with guidance of multimodal reference, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 05, 2020, pp. 9749–9756, http://dx.doi.org/10.1609/aaai.v34i05.6525, URL https://ojs.aaai.org/index.php/AAAI/article/view/6525.
[25] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, H. Wang, ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 4, 2021, pp. 3208–3216.
[26] X. Hu, X. Yin, K. Lin, L. Zhang, J. Gao, L. Wang, Z. Liu, Vivo: Visual vocabulary pre-training for novel object captioning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 2, 2021, pp. 1575–1583.
[27] L. Sun, J. Wang, K. Zhang, Y. Su, F. Weng, RpBERT: A text-image relation propagation-based BERT model for multimodal NER, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, no. 15, 2021, pp. 13860–13868.
[28] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, R. Huang, Sarcasm as contrast between a positive sentiment and negative situation, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 704–714.
[29] A. Ghosh, T. Veale, Fracking sarcasm using neural network, in: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2016, pp. 161–169.
[30] C. Baziotis, N. Athanasiou, P. Papalampidi, A. Kolovou, G. Paraskevopoulos, N. Ellinas, A. Potamianos, NTUA-SLP at SemEval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive RNNs, 2018, arXiv preprint arXiv:1804.06659.
[31] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intell. Syst. 34 (3) (2019) 38–43.
[32] R. Schifanella, P. De Juan, J. Tetreault, L. Cao, Detecting sarcasm in multimodal social platforms, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 1136–1145.
[33] D. Tomás, R. Ortega-Bueno, G. Zhang, P. Rosso, R. Schifanella, Transformer-based models for multimodal irony detection, J. Ambient Intell. Humaniz. Comput. (2022) 1–12.
[34] M. Malik, D. Tomás, P. Rosso, How challenging is multimodal irony detection? in: International Conference on Applications of Natural Language to Information Systems, Springer, 2023, pp. 18–32.
[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.
[36] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for MobileNetV3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.
[37] K. He, Y. Huang, R. Mao, T. Gong, C. Li, E. Cambria, Virtual prompt pre-training for prototype-based few-shot relation extraction, Expert Syst. Appl. (ISSN: 0957-4174) 213 (2023) 118927, http://dx.doi.org/10.1016/j.eswa.2022.118927.

[38] K. Maity, P. Jha, S. Saha, P. Bhattacharyya, A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1739–1749.
[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022.
[42] Y. Kim, Convolutional neural networks for sentence classification, 2014, Eprint arxiv, arXiv:1408.5882.
[43] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5–6) (2005) 602–610.
[44] Y. Tay, L.A. Tuan, S.C. Hui, J. Su, Reasoning with sarcasm by reading in-between, 2018, arXiv preprint arXiv:1805.02856.
[45] T. Xiong, P. Zhang, H. Zhu, Y. Yang, Sarcasm detection with self-matching networks and low-rank bilinear pooling, in: The World Wide Web Conference, 2019, pp. 2115–2124.
[46] J.D.M.-W.C. Kenton, L.K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.
[47] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, VisualBERT: A simple and performant baseline for vision and language, 2019, arXiv preprint arXiv:1908.03557.
[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017, arXiv:1412.6980.
[49] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, S. Lee, 12-In-1: Multi-task vision and language representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10437–10446.
[50] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[51] D.Q. Nguyen, T. Vu, A.T. Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 9–14.
[52] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, 2020, arXiv:1909.11942.
[53] R. Mao, X. Li, K. He, M. Ge, E. Cambria, MetaPro Online: A computational metaphor processing online system, in: Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics, ACL Demonstration Track, 2023.