Modeling Intra and Inter-modality Incongruity for Multi-Modal Sarcasm Detection

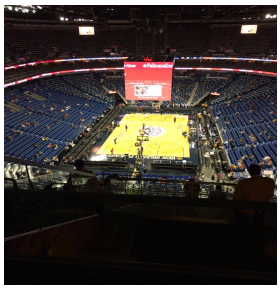
Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, Weiping Wang
Chinese Academy of Sciences Bejing, China
Institute of Information Engineering Bejing, China
{panhongliang,linzheng,fupeng,qiyatao,wangweiping}@iie.ac.cn

Abstract

Sarcasm is a pervasive phenomenon in today's social media platforms such as Twitter and Reddit. These platforms allow users to create multi-modal messages, including texts, images, and videos. Existing multi-modal sarcasm detection methods either simply concatenate the features from multi modalities or fuse the multi modalities information in a designed manner. However, they ignore the incongruity character in sarcastic utterance, which is often manifested between modalities or within modalities. Inspired by this, we propose a BERT architecture-based model, which concentrates on both intra and inter-modality incongruity for multi-modal sarcasm detection. To be specific, we are inspired by the idea of self-attention mechanism and design inter-modality attention to capturing inter-modality incongruity. In addition, the co-attention mechanism is applied to model the contradiction within the text. The incongruity information is then used for prediction. The experimental results demonstrate that our model achieves state-of-the-art performance on a public multi-modal sarcasm detection dataset.

1 Introduction

Sarcasm is a form of figurative language where the literal meaning of words does not hold, and instead, the opposite interpretation is intended (Joshi et al., 2017). Sarcasm is prevalent in today's social media platforms, and it can completely flip the polarity of sentiment or opinion. Thus, an effective sarcasm detector is beneficial to applications like sentiment analysis, opinion mining (Pang and Lee, 2007), and other tasks that require people's real sentiment. However, the figurative nature of sarcasm makes it a challenging task (Liu, 2010). The scholars notice that sarcasm is often associated with a concept called incongruity which is used to suggest a distinction between reality and expectation (Gibbs Jr et al., 1994). Consequently, many approaches for sarcasm detection have been proposed by capturing the incongruity within text (Riloff et al., 2013; Joshi et al., 2015; Tay et al., 2018; Xiong et al., 2019).



(a). such a packed game . it is amazing we even got a seat . # pelicans

(b). well that looks appetising ... # ubereats

Figure 1: Examples of image modality aiding sarcasm detection. (a) It suggests a contradiction of "it is amazing we even got a seat" in the text and "many unoccupied seats" on the image. (b) The food on the image doesn't look appetising as the text describes.

More and more applications like Twitter allow users to post multi-modal messages. Accordingly, only modeling the incongruity within text modality is not enough to identify the inter-modality contradiction's sarcasm. Consider the given examples in Figure 1; people can not recognize sarcasm merely from text unless they find the contradiction between text and images. As a result, capturing the incongruity between modalities is significant for multi-modal sarcasm detection.

However, the existing models for multi-modal sarcasm detection either concatenate the features from multi modalities (Schifanella et al., 2016) or fuse the information from different modalities in a designed manner (Cai et al., 2019). Previous multi-modal sarcasm detection approaches neglect the incongruity character of sarcasm. We believe that it is

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1383–1392

meaningful to capture both intra and inter-modality incongruity for multi-modal sarcasm detection.

We treat images and text as two modalities in this work and propose a novel BERT architecture-based model for multi-modal sarcasm detection. BERT as a pre-trained language model proposed by Devlin et al. (2019), which can be used to produce outstanding representations of text. For this reason, we utilize BERT to acquire the representation of text and the hashtags (use the word with a '#' in front to indicate the topic of the tweet) within the text. We notice that hashtags might contain the information that contrasts the text. Maynard and Greenwood (2014) also studies the sentiment and sarcasm with the help of hashtags. Consequently, we apply a co-attention matrix to model the incongruity between text and hashtags as the intra-modality incongruity. Besides, the self-attention mechanism considers the interaction between keys and queries and the inter-modality incongruity information can also be treated as an interaction between text and images. As a result, inspired by the key idea of self-attention, we design the inter-modality attention which treats textual features as queries, image features as keys and values to capture the inter-modality incongruity. The intra and inter-modality incongruity information are then combined for prediction.

The main contributions of our work can be summarised as follows:

- We propose a novel BERT architecture-based model for multi-modal sarcasm detection, aiming to address the problem that existing multi-modal sarcasm detection models do not consider the incongruity character of sarcasm.
- We design the inter-modality attention to model the incongruity between modalities and apply the co-attention mechanism to model the incongruity within text modality for multi-modal sarcasm detection.
- We conduct a series of experiments to show our model's effectiveness and our model achieves a 2.74% improvement on F1 score than state-of-the-art method. Furthermore, we find that considering the text on the images can bring significant improvements.

2 Method

In this section, we first define the multi-modal sarcasm detection task. We then briefly present the background of the BERT model and describe the architecture of our proposed model in detail. Figure 2 gives an overview of our model.

2.1 Task Definition

Multi-modal sarcasm detection aims to identify if a given text associated with an image has sarcastic meaning. Formally, given a set of multimodal samples D, for each sample $d \in D$, it contains a sentence T with n words $\{ t_1, t_2, t_3, \ldots, t_n\}$ and an associated image I. The goal of our model is to learn a multi-modal sarcasm detection classifier to correctly predict the results of unseen samples.

2.2 Background

Language model pretraining has been proven to be useful for many natural language processing tasks (Peters et al., 2018; Howard and Ruder, 2018). BERT was proposed by Devlin et al. (2019), which is designed to pre-train deep bidirectional representations from large unlabelled data by jointly conditioning on both left and right context in all layers. The pretraining procedure makes BERT have the capacity to acquire well representations of text. The BERT model consists of multi-layer bi-directional transformer encoders (Vaswani et al., 2017). Devlin et al. (2019) propose two BERT models in their work. A Base BERT model with 12 transformer blocks, feed-forward networks with 768 hidden units and 12 attention heads, and a Large BERT model with 24 transformer blocks, feed-forward networks with 1024 hidden units and 16 attention heads, In our work, we apply a pre-trained Base BERT model to obtain text representations.

2.3 Model Architecture

Our model can be divided into three parts: the Image and Text Processing module, the inter-modality attention module, and the intra-modality attention module.

Image and Text Processing

For text processing, given a sequence of words $X = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in \mathbb{R}^d$ is the sum-up of word, segment, and position embeddings, N is the maximum length of the sequence and d is the embedding size. We adopt the pre-trained BERT model on it to acquire text representations. The encoded text can be depicted as $H \in \mathbb{R}^{d*N}$, which is the output of the last layer of BERT encoders and d is the hidden size of BERT.

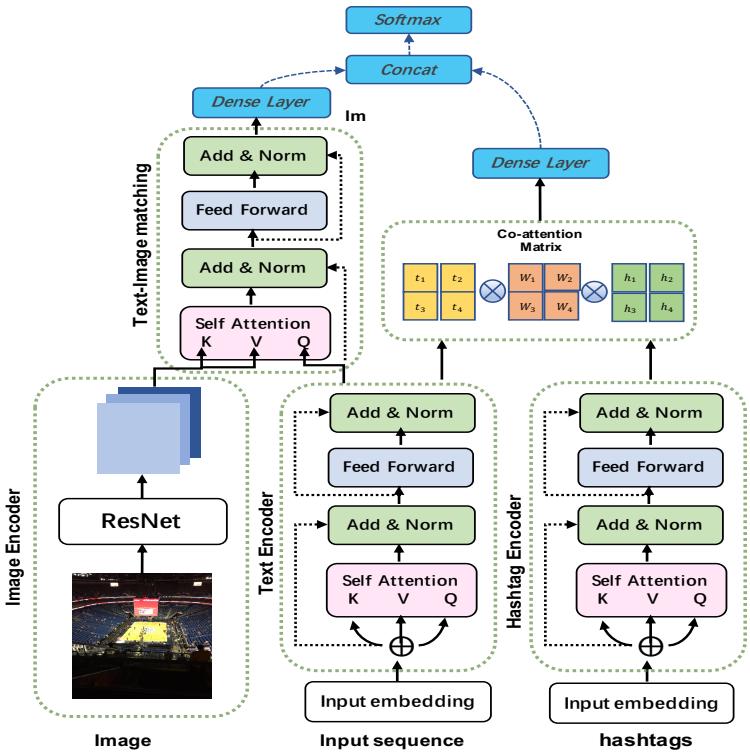


Figure 2: Overview of our proposed model. A pre-trained BERT model encodes a given sequence and the hashtags within it. ResNet is used to obtain the image representation. We apply intra-modality attention to model the incongruity within the text and inner-modality attention to model the incongruity between text and images. The incongruity information is then combined and used to predict.

As for image processing, given an image I, we first resize it to 224*224 pixels, and then we use ResNet-152 (He et al., 2016) to obtain the representation of the image. To be specific, we chop off the last fully-connected (FC) layer and obtain the output of the last convolutional layer:

$$ResNet(I) = \{r_i | r_i \in \mathbb{R}^{2048}, i = 1, 2, \ldots, 49\} \quad (1)$$

where each r_i is a 2048-dimensional vector representing a region on the image. Consequently, an image I can be represented as $ResNet(I) \in \mathbb{R}^{2048*49}$. Finally, in order to project the visual features into the same dimension of textual features, we conduct a linear transformation on the encoded image representation $ResNet(I)$ as:

$$G = W_v ResNet(I) \quad (2)$$

where $W_v \in \mathbb{R}^{d*2048}$ is a trainable parameter and d is the diemnsion of textual feature. $G \in \mathbb{R}^{d*49}$ is the ecoded representation of visual features.

Inter-modality Attention

Self-attention can be used to generate an internal representation of a sequence. The internal representation considers the interaction between each pair of tokens in the sequence. Inter-modality incongruity information can be represented as a kind of interaction between the features of multi modalities. Particularly, the input tokens will give high attention values to the image regions contradicting them as incongruity is a key character of sarcasm. Hence, we borrow the idea from the self-attention mechanism and design a text-image matching layer to capture the incongruity information between text and images. Our text-image matching layer accepts the text features $H \in \mathbb{R}^{d*N}$ as queries, and the image features $G \in \mathbb{R}^{d*49}$ as keys and values. In this way, the text features can guide the model to pay more attention to the incongruous image regions. Specifically, for the ith head of the text-image matching layer, it has the following form:

$$ATT_i(H, G) = softmax(\frac{[W_i^Q H]^T [W_i^K G]}{\sqrt{d_k}})[W_i^V G]^T \quad (3)$$

where $d_k \in \mathbb{R}^{d/h}$, $ATT_i(H, G) \in \mathbb{R}^{N*d_k}$, and $\{W_i^Q, W_i^K, W_i^V\} \in \mathbb{R}^{d_k*d}$ are learnable parameters. The outputs of h heads are then concatenated and followed by a linear transformation as:

$$MATT(H, G) = [ATT_1(H, G), \ldots, ATT_h(H, G)]W^o \quad (4)$$

where $W^o \in \mathbb{R}^{d*d}$ is a learnable parameter. After that, a residual connection is worked on the text feature H and the output of self-attention layer $MATT(H, G)$ as:

$$Z = LN(H + MATT(H, G)) \tag{5}$$

where LN is the layer normalization operation proposed by Ba et al. (2016). After that, a feed-forward network (a.k.a MLP) and another residual connection are employed on Z to obtain the output of the first transformer encoder:

$$TIM(H, G) = LN(Z + MLP(Z)) \tag{6}$$

where $TIM(H, G) \in \mathbb{R}^{N*d}$ is the output of the first text-image matching layer. We stack l_m such text-image matching layers and get $TIM_{l_m}(H, G)$ as the output of the last layer, where $TIM_{l_m}(H, G) \in \mathbb{R}^{N*d}$ and l_m is a pre-defined hyper-parameter. The final representation of inter-modality incongruity can be describes as $H_G \in \mathbb{R}^d$, which is the encoding of $[CLS]$ token in $TIM_{l_m}(H, G)$

Intra-modality Attention

As the incongruity might only appear within the text (e.g., a sarcastic text associated with an unrelated image), it is necessary to consider the intra-modality incongruity. Social media like Twitter allow users to add hashtags to indicate the topic or their real minds. Maynard and Greenwood (2014) point out that hashtags are useful when analyzing a user's real sentiment (e.g., I am happy that I woke up at 5:15 this morning. # not). Accordingly, we take the contradiction between the original text and the hashtags within it as intra-modality incongruity (i.e., for those samples without hashtags, we use a special token instead). Intuitively, we can use the same way as inter-modality attention to gain the intra-modality incongruity information. However, we find that it doesn't bring much improvement even it contains more parameters. Hence, inspired by Lu et al. (2016) 's work, we introduce an affinity matrix C to model the interaction between the text and the hashtags. C is calculated by:

$$C = tanh(H^T W_b T) \tag{7}$$

where $H \in \mathbb{R}^{d*N}$ and $T \in \mathbb{R}^{d*M}$ represent the text features and the hashtag features separately. N and M are pre-defined hyper-parameters denoting the input sequence's max length and hashtags, respectively. $W_b \in \mathbb{R}^{d*d}$ is a learnable parameter containing weights. After computing the affinity matrix $C \in \mathbb{R}^{N*M}$, we maximize the affinity matrix over text features' locations to get hashtag attention. To be specific, we compute a weight vector $a \in \mathbb{R}^M$ by applying a column-wised max-pooling operation on the matrix C. Tay et al. (2018) argues that the words that contribute to the incongruity (usually accompany with a high attention value) should be highlighted. Therefore, a more discriminative pooling operator like max-pooling is desirable in our case. Finally, the intra-modality incongruity is computed as:

$$H_T = aT^T \tag{8}$$

where $H_T \in \mathbb{R}^d$ contains the intra-modality incongruity information.

2.4 Prediction

After obtaining the intra-modality incongruity representation H_T and inter-modality incongruity representation H_G, we concatenate them for prediction. The prediction part consists of a linear layer to reduce the dimension and a $Softmax$ function to distribute probabilities to each category. Our model will classify the given text into the category with the highest probability. This procedure can be described as:

$$\hat{y} = Softmax(W[H_G : H_T] + b) \tag{9}$$

where $W \in \mathbb{R}^{2d}$ is learnable parameter training along with the model. $\hat{y}$ is the classification result of our model.

2.5 Training objectives

Cross-entropy loss function is used in our work for optimizing the model.

$$J = -\sum_{i=1}^{N} [y_i log\hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] + \lambda R \tag{10}$$

where J is the cost function. $\hat{y}_i$ is the prediction result of our model for sample i, and y_i is the true label for sample i. N is the size of training data. R is the standard L2 regularization and λ is the weight of R.

3 Experiment

This section first describes the dataset, experimental settings, baseline models, and experimental results. Then, we conduct a series of ablative experiments to verify the components' effectiveness in

	Training	Development	Testing
Sentences	19816	2410	2409
Positive	8642	959	959
Negative	11174	1451	1450
Avg length	15.71	15.72	15.89

Table 1: Dataset description

our model. After that, we analyze the influence of the number of text-image matching layers on model performance. Finally, we give out a model visualization on several given sarcastic cases and perform an analysis of the wrongly predicted samples.

3.1 Dataset

We evaluate our model on a publicly available multi-modal sarcasm detection dataset,[1] which is collected by Cai et al. (2019). Each sample in the dataset consists of a sequence of text and an associated image. The tweets containing the words like $sarcasm$, $sarcastic$, $irony$, $ironic$ or $URLs$ are discarded during data pre-processing. Cai et al. (2019) divides the data into a training set, a development set, and a testing set with a ratio of 80%:10%:10%. They also manually check the development set and testing set to ensure the accuracy of labels. Detailed statistics are summarized in Table 1.

3.2 Baseline Models

We divide the baseline models into three categories: visual modality models, Text modality models, and Text+Visual modality models.

- **Visual modality models:**
  Image-Only: The image feature G is directly used to predict the results after an average pooling operation.

- **Text modality models:**
  TextCNN: It is proposed by Kim (2014), which is a deep learning model based on CNN for addressing text classification tasks.
  SIARN: SIARN is proposed by Tay et al. (2018). It employs inner-attention for textual sarcasm detection to overcome the weakness of previous sequential models such as RNNs, which cannot capture the interaction between word pairs and hampers the ability to explicitly model incongruity.
  SMSD: Following the work of (Tay et al., 2018), Xiong et al. (2019) propose a self-matching network to capture sentence incongruity information by exploring word-to-word interaction.
  BERT: BERT as a pre-trained model proposed by Devlin et al. (2019) , which achieves state-of-the-art results in many NLP tasks. We consider it a baseline to investigate whether the performance gain comes from BERT or our proposed method.

- **Visual+Text modality models:**
  Hierarchical Fusion Model(HFM): Cai et al. (2019) propose a Hierarchical Fusion Model for multi-modal sarcasm detection. Their model takes image features, image attribute features, and text features as three modalities. Features of three modalities are reconstructed and fused for prediction.
  Res-bert: We implement Res-bert as one of our baseline models. Res-bert simply concatenates the image features G, and text feature H for classification.

Hyper-parameters	Value
Batch size	32
Learning rate	5e-5
Weight decay	1e-2
Epochs	8
Gradient clipping	1.0
Warmup rate	0.2
Text length	75
Hashtag length	10
Dropout rate	0.1

Table 2: Hyper-parameters

3.3 Experimental Settings

Our model is implemented in PyTorch (Paszke et al., 2019), running on a NVIDIA TITAN RTX GPU. The pre-trained BERT model is available from the $Transformers$ toolkit released by Hugging Face.[2] We adopt Adam (Kingma and Ba, 2015) as our optimizer and set the initial learning rate as 5e-5 with a warmup rate of 0.2. The batch size is fixed to 32 for training. The maximum length is 75 for text and 10 for hashtags, respectively. Our model is fine-tuned for eight epochs on

[1] https://github.com/headacheboy/data-of-multimodal-sarcasm-detection

[2] https://huggingface.co/transformers/

Modality	Method	Precision	Recall	Accuracy	F1 score
	Random	0.4055	0.5057	0.5027	0.4470
Visual	Image-Only	0.6511	0.6715	0.7260	0.6611
Text	TextCNN	0.7429	0.7639	0.8003	0.7532
	SIARN	0.7555	0.7570	0.8057	0.7563
	SMSD	0.7646	0.7518	0.8090	0.7582
	BERT	0.7827	0.8227	0.8385	0.8022
Visual+Text	HFM	0.7657	0.8415	0.8344	0.8018
	Res-bert	0.7887	0.8446	0.8480	0.8157
	Method (this paper)	**0.8087**	**0.8508**	**0.8605**	**0.8292**

Table 3: Experiment results on the multi-modal sarcasm detection dataset. The best results are in bold.

the training set. We save the model, which has the best performance on the validation set. The full parameters are listed in Table 2.

3.4 Experimental Results

We compare our model with the baseline models on the standard metrics, including precision, recall, F1 score, and accuracy.[3] The results are shown in Table 3. The experimental results illustrate that our model achieves the best performance across the baseline models. Specifically, our model obtains a 2.74% improvement in terms of F1 score compared with the state-of-the-art Hierarchical Fusion Model (HFM) proposed by Cai et al. (2019). Our model also outperforms the fine-tuned BERT model with a 2.7% improvement, which shows our model's effectiveness and the important role of the images.

We can see from table 3, the model only using image features does not perform well, which demonstrates that images cannot be treated independently for the multi-modal sarcasm detection task. Obviously, the methods based on text modality achieve better performance than the method based on image modality. Consequently, text information is more useful than image information for sarcasm detection. It is worth noticing that the fine-tuned BERT model performs far better than other text-based non-pre-trained models, which supports our motivation that pre-trained models like BERT can improve our task. The models belonging to Visual+Text modality generally achieve better results than the others, indicating that images are useful to enhance performance.

Looking at the models inside text modality, both SIARN (Tay et al., 2018) and SMSD (Xiong et al., 2019) take incongruity information into consideration and outperform TextCNN. Hence, the incongruity information is beneficial to identify sarcasm. Our proposed method achieves better results than Res-bert, proving that modeling both intra and inter-modality incongruity is more effective than a simple concatenation of modalities for multi-modal sarcasm detection.

[3]We implement the metrics by using sklearn.metrics.

Model	Precision	Recall	Acc	F1
BERT	0.7827	0.8227	0.8385	0.8022
Model(w\o inter)	0.7764	0.8508	0.8430	0.8119
Model(w\o intra)	0.8005	0.8373	0.8522	0.8185
Method (this paper)	**0.8087**	**0.8508**	**0.8605**	**0.8292**

Table 4: Ablation experiment results. The best results are in bold.

3.5 Ablation Study

To evaluate the effectiveness of the components in our model, we conduct a series of ablative experiments. We first remove the intra-modality attention and get model(w\o intra), which only uses H_G for prediction. Then, we eliminate the inter-modality attention and get model(w\o inter). This model concatenates H and H_T to the classifier layer as the experimental results indicate that H_T only plays a supporting role in our model.

Table 4 gives the results of ablative experiments. It shows that our proposed model achieves the best performance when including both intra and inter-modality attention modules. The absence of inter-modality attention leads to decreased results, proving that considering the contradiction between modalities is meaningful for multi-modal sarcasm detection. The model without the intra-modality attention also impedes the performance. As a result, both intra and inter-modality attention plays an indispensable role in our model.

(a). such a packed game . it is amazing we even got a seat . # pelicans

(b). well that looks appetising ... ubereats

(c). good thing my 2nd graders are not distracted by chainsaws , falling trees , and chippers

Figure 3: The figure illustrates the attention visualization of some sarcastic tweets. We find our model is capable of focusing its attention on the incongruous regions, marked by bright colour.

3.6 Model Analysis

The impact of the number of text-image matching layer l_m:

We measure the model performance on the F1 score along with a range of the text-image matching layer number lm from 1 to 7. We can see in Figure 4, the F1 score increases until reaching a peak point when l_m equals to 3. Our model achieves the best performance at this point. Then, the model performance begins to decrease as l_m continues to grow. We guess the performance worsens, probably due to the increase of the model parameter, suggesting that adding more text-image matching layers might not enhance but impede the performance.

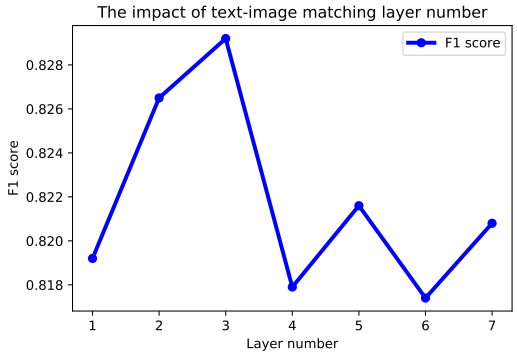


Figure 4: The performance curves with a variety of l_m from 1 to 7.

Model visualization:

In this section, we visualize the text-image attention distributions. Our model is designed to capture the incongruity information. Therefore, incongruous regions on the images are more likely to be attended by our model. We demonstrate several sarcastic cases collected from the dataset:

- "such a packed game . it's amazing we even got a seat . # pelicans"
- "well that looks appetising ... # ubereats"
- "good thing my 2nd graders aren not distracted by chainsaws , falling trees , and chippers !"

Figure 3 illustrates that our model is highly effective in attending the incongruous regions. In the first example, our model attends to the regions indicating "lots of unoccupied seats," which forms a contradiction with the text "it is amazing we even got a seat.". Similar patterns can also be noticed in the second and third instances.

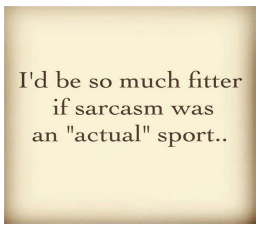
Model	Precision	Recall	Acc	F1
Method (this paper)	0.8087	0.8508	0.8605	0.8292
Method (adding text)	**0.8433**	**0.8811**	**0.8875**	**0.8618**

Table 5: Experiment results when involving the text on the image in our model.

Error analysis:

We also perform a qualitative analysis of the wrongly predicted samples. We check approximately 50 false classified instances and find that our model might incorrectly classify those samples containing necessary text information on the images (see Figure 5). Consequently, considering the text on the images might bring improvements for the multi-modal sarcasm detection task. Based on this observation, we further implement an experiment in which the text on the images is considered. Specifically, we apply a General Character Recognition API to acquire the text on the pictures and use a co-attention matrix to model the incongruity information between the original tweet and the text. Table 5 shows that our model achieves a significant improvement when considering the text on the images. In addition, we find that our model

might struggle in those instances requiring external knowledge, such as a speaker's facial gesture or contextual information. Thus, external information is also essential for sarcasm detection.



(a). I could enter the Olympics !

(b). Inspiring quote for the day .

Figure 5: Wrongly classified samples with important textual information on the image.

4 Related Work

4.1 Text-based Sarcasm detection

The existing text-based approaches can be classified into three categories: rule-based approaches, feature-based machine learning approaches, and deep learning-based approaches (Joshi et al., 2017). Rule-based methods aim to spot sarcasm by detecting some fixed patterns. Riloff et al. (2013) observe that a common form of sarcasm that both positive sentiment and negative situation appear simultaneously. Inspired by this, they develop a bootstrapping algorithm that iteratively expands positive and negative phrase sets. The learned phrases are then used to detect sarcasm. Maynard and Greenwood (2014) design a hashtag tokenizer to analyze the sentiment and sarcasm within hashtags. They also compile a set of rules to determine the sentiment polarity when knowing sarcasm.

However, rule-based methods strongly rely on the collected patterns, and it is challenging to identify the sarcasm caused by uncollected patterns. Accordingly, researchers begin to design various textual features and apply machine learning methods for recognizing sarcasm. Joshi et al. (2015) develop a system considering lexical features, pragmatic features, and incongruity features. SVM is used as their classifier. Ghosh et al. (2015) also apply SVM as their classifier and treat sarcasm detection as a word sense disambiguation problem.

Though machine learning approaches have achieved significant improvement, feature extraction is a time-consuming job. Recent works are mainly based on deep learning methods as they are capable of automatically extracting features and obtain promising results. Poria et al. (2016) use pre-trained CNNs to extract sentiment, emotion and personality features for sarcasm detection. Both Tay et al. (2018) and Xiong et al. (2019) try to explicitly model the incongruity between the word pairs using attention mechanism and receive satisfying results.

4.2 Multi-modal Sarcasm detection

It is worth noticing that there are also some valuable works concentrating on multi-modal sarcasm detection. Schifanella et al. (2016) first consider both textual and visual features for sarcasm detection and propose two alternative frameworks. Mishra et al. (2017) propose a cognitive NLP system for sentiment and sarcasm classification. They introduce a framework to extract cognitive features from the eye-movement/gaze data automatically. They use CNN to encode both gaze-based and textual features for classification. Castro et al. (2019) propose a new sarcasm dataset, compiled from TV shows. They treat text features, speech features, and video features as three modalities and use SVM as the classifier. Cai et al. (2019) introduce a hierarchical fusion model. They take image features, image attribute features, and text features as three modalities. Features of three modalities are reconstructed and fused for prediction.

5 Conclusion

In this paper, we propose a novel BERT architecture-based model to address the issue that existing multi-modal sarcasm detection approaches do not consider incongruity character of sarcasm. To be specific, our model considers both intra and inter-modality incongruity and achieves state-of-the-art performance on a public multi-modal sarcasm detection dataset. Besides, we also conduct a series of experiments to verify the effectiveness of our model. Finally, we perform error analysis and find that the text on the images is essential for multi-modal sarcasm detection.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Conference of the Association for Computational Linguis-*

tics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 2506–2515.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 4619–4629.

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.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words. In *EMNLP 2015*, pages 1003–1012.

Raymond W Gibbs Jr, Raymond W Gibbs, and Jr Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5):73:1–73:22.

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 ACL 2015*, pages 757–762.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition.*

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 289–297.

Diana Maynard and Mark A. Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 4238–4243.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *ACL 2017*, pages 377–387.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 8024–8035.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *COLING 2016*, pages 1601–1612.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 704–714.

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.

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, ACL 2018*, pages 1010–1020.

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, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

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*, pages 2115–2124.