Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association

Nan Xu, Zhixiong Zeng, Wenji Mao
Institute of Automation, Chinese Academy of Sciences
School of Artificial Intelligence, University of Chinese Academy of Sciences
{xunan2015,zengzhixiong2018,wenji.mao}@ia.ac.cn

Abstract

Sarcasm is a sophisticated linguistic phenomenon to express the opposite of what one really means. With the rapid growth of social media, multimodal sarcastic tweets are widely posted on various social platforms. In multimodal context, sarcasm is no longer a pure linguistic phenomenon, and due to the nature of social media short text, the opposite is more often manifested via cross-modality expressions. Thus traditional text-based methods are insufficient to detect multimodal sarcasm. To reason with multimodal sarcastic tweets, in this paper, we propose a novel method for modeling cross-modality contrast in the associated context. Our method models both cross-modality contrast and semantic association by constructing the Decomposition and Relation Network (namely D&R Net). The decomposition network represents the commonality and discrepancy between image and text, and the relation network models the semantic association in cross-modality context. Experimental results on a public dataset demonstrate the effectiveness of our model in multimodal sarcasm detection.

1 Introduction

Sarcasm is a sophisticated linguistic phenomenon, defined by Merriam-Webster Dictionary as 'The use of words that mean the opposite of what you really want to say, especially in order to insult someone, to show irritation, or to be funny'. It can not only disguise the hostility of the speaker, but also enhance the effect of mockery or humor on the listener (Tay et al., 2018). As an important clue to analyze people's true sentiment and intentions in communication from implicit expressions, automatic sarcasm detection plays a significant role in various applications that require the knowledge of people's sentiment or opinion (Cai et al., 2019), such as customer service, political stance detection, and user intent recognition.

Figure 1: Examples of multimodal tweets. The non-sarcasm (a) shows the user's affection for the beautiful trees with positive sentiment; and (b) is a sarcastic tweet where the text word 'perfect' contrasts sharply with the rainy weather in the image

Existing work on sarcasm detection mainly focuses on text data. Early feature engineering approaches rely on the signal indicators of sarcasm, such as syntactic patterns, lexical indicators and special symbols (Tsur et al., 2010; Davidov et al., 2010; González-Ibánez et al., 2011). As sarcasm is often associated with implicit contrast or disparity between conveyed sentiment and user's situation in context (Riloff et al., 2013), contextual contrast information at conversation, tweet or word level is also employed to detect sarcasm in text (Bamman and Smith, 2015; Rajadesingan et al., 2015; Joshi et al., 2016). Recently, deep learning based methods are adopted to train end-to-end neural networks (Baziotis et al., 2018; Tay et al., 2018), achieving state-of-the-art performance.

With the fast growing and diverse trend of social media, multimodal sarcastic tweets which convey abundant user sentiment are widely posted on various social platforms. There is a great demand for multimodal sarcasm detection to facilitate various applications. However, traditional text-based methods are not applicable to detect multimodal sarcastic tweets (Fig.1). In multimodal context, sarcasm is no longer a pure linguistic phenomenon,

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3777–3786
July 5 - 10, 2020.

but rather the combined expressions of multiple modalities (i.e. text, image, etc.). As the short text in tweet often has insufficient contextual information, contextual contrast implied in multimodal sarcasm is typically conveyed by cross-modality expressions. For example, in Fig.1b, we can not reason about sarcasm intention simply from the short text 'Perfect flying weather in April' until we notice the downpour outside the airplane window in the attached image. Therefore, compared to text-based methods, the essential research issue in multimodal sarcasm detection is the reasoning of cross-modality contrast in the associated situation.

Several related work on multimodal sarcasm detection has been proposed (Schifanella et al., 2016; Cai et al., 2019; Castro et al., 2019). However, they mainly focus on the fusion of multimodal data, and did not address the above key research issue in reasoning with multimodal sarcasm. There are still two main research challenges for multimodal sarcasm detection. First, since sarcasm commonly manifests with a contrastive theme, this requires the detection model to have the ability to reason about cross-modality contrast or incongruity of situations. Second, to ensure cross-modality contrast assessed in the associated common ground, the detection model needs the mechanism to concentrate on the semantic associated aspects of situations in cross-modality context. This contextual contrast and semantic association information acquired, in turn, can provide salient evidence to interpret the detection of multimodal sarcasm.

To tackle the above challenges, in this paper, we propose a novel method to model both cross-modality contrast and semantic association by constructing the Decomposition and Relation Network (i.e. D&R Net) for multimodal sarcasm detection task. The decomposition network implicitly models cross-modality contrast information via representing the commonality and discrepancy between image and text in tweets. The relation network explicitly captures the semantic association between image and text via a cross-modality attention mechanism. The main contributions of our work are as follows:

- We identify the essential research issue in multimodal sarcasm detection, and propose a method to model cross-modality contrast in the associated context of multimodal sarcastic tweets.
- We construct the Decomposition and Relation Network (D&R Net) to implicitly represent the contextual contrast and explicitly capture the semantic association between image and text, which provides the reasoning ability and word-level interpretability for multimodal sarcasm detection.
- We compare our model with the existing state-of-the-art methods, and experimental results on a publicly available dataset demonstrate the effectiveness of our model in multimodal sarcasm detection.

2 Related Work

2.1 Textual Sarcasm Detection

Traditional sarcasm detection takes text-based approaches, including feature engineering, context based and neural network models. Earlier feature engineering approaches are based on the insight that sarcasm usually occurs with specific signals, such as syntactic patterns (e.g. using high-frequency words and content words) (Tsur et al., 2010), lexical indicators (e.g. interjections and intensifiers) (González-Ibánez et al., 2011), or special symbols (e.g. '?', '!', hashtags and emojis) (Davidov et al., 2010; Felbo et al., 2017). As sarcasm is often associated with an implicit contrast or disparity between conveyed sentiment and user's situation in context (Riloff et al., 2013), some studies rely on this basic character of sarcasm to detect contextual contrast at different linguistic levels, including immediate communicative context between speaker and audience (Bamman and Smith, 2015), historical context between current and past tweets (Rajadesingan et al., 2015; Joshi et al., 2015), or word-level context by computing semantic similarity (Hernández-Farías et al., 2015; Joshi et al., 2016).

Recently, researchers utilize the powerful techniques of neural networks to get more precise semantic representations of sarcastic text and model the sequential information of sarcastic context. Some approaches consider the contextual tweets of target tweet, using RNN model for contextual tweets representation and modeling the relationship between target and contextual tweets for sarcastic text classification (González-Ibánez et al., 2011; Zhang et al., 2016). To conceive more indicative information, user embedding (Amir et al., 2016), emotion, sentiment, personality (Poria et al., 2016), speaker's psychological profile (Ghosh and Veale,

2017), cognitive features (Mishra et al., 2017), and syntactic features (Baziotis et al., 2018) are also incorporated into CNN/LSTM models to enhance the performance. Furthermore, to overcome the black box problem of neural network model and reasoning with sarcasm, some novel methods such as neural machine translation framework (Peled and Reichart, 2017), and intra-attention mechanism (Tay et al., 2018) are explored to improve the interpretability of sarcasm detection.

2.2 Multimodal Sarcasm Detection

With the prevalence of multimodal tweets, multimodal sarcasm detection has gained increasing research attention recently. Schifanella et al. (2016) firstly tackle this task as a multimodal classification problem and concatenate manually designed features of image and text to classify sarcasm. Cai et al. (2019) extend the input modalities with triple features (i.e. text feature, image feature and image attributes), and propose a hierarchical fusion model for the task. Castro et al. (2019) firstly propose video-level multimodal sarcasm detection task and deal with it based on feature engineering via SVM. However, these methods pay more attention to the fusion of multimodal features, and did not consider cross-modality contrast and semantic association information which is essential to deduce multimodal sarcastic tweets.

In this paper, we propose a novel method to model the cross-modality contrast and semantic association in multimodal context by constructing the Decomposition and Relation Network (D&R Net), which enables our model to reason with multimodal sarcastic tweets and provides pertinent evidence for interpretation.

3 Proposed Model

Fig.2 illustrates the overall architecture of our proposed D&R Net for multimodal sarcasm detection, which is composed of four modules, preprocessing, encoding, decomposition network and relation network. We first preprocess the image and text inputs and extract adjective-noun pairs (ANPs) from each image. We then encode these triple inputs into hidden representations. After that, we learn to represent the commonality and discrepancy between image and text in decomposition network as well as the multi-view semantic association information in relation network. Finally, we feed these cross-modality representations into classification module for multimodal sarcasm detection.

3.1 Preprocessing

Standard image, text and visual attributes (e.g. sunnet, scene, snow) are utilized in the previous multimodal sarcasm detection (Cai et al., 2019). To enhance the image semantic understanding, we practice a better way to get more visual semantic information via extracting extra adjective-noun pairs from each image (e.g. great sunset, pretty scene, fresh snow in Fig.2). Thus, our model accepts triple inputs.

$$Input = [Text, Image, ANPs] \quad (1)$$

where, $Text = [W_j]_j^T$, T is the length of text sequence; $ANPs = [P_i]_i^N$, N is the number of adjective-noun pair, in which each pair P_i contains an adjective word A_i, a noun word N_i and the probability value p_i of this kind of ANP existing in the attached $Image$, $P_i = [(A_i, N_i), p_i]$.

3.2 Encoding

In encoding module, we map these triple inputs into hidden representations. All textual words W_j, A_i, N_i are firstly mapped into embedding vectors $w_j, a_i, n_i \in \mathbb{R}^d$.

For each text, we utilize the bi-directional long short term memory (BiLSTM) network to represent textual sequence into a hidden representation vector and incorporate the contextual information. It maps word embedding w_j into hidden state $h_j^w \in \mathbb{R}^d$.

$$H^w = [h_j^w] = BiLSTM([w_j]) \in \mathbb{R}^{T \times d} \quad (2)$$

For each ANP, we directly compute the maxpooling result of its adjective and noun word embeddings as the hidden representation.

$$H^p = [h_i^p] = MaxPooling([a_i, n_i]) \in \mathbb{R}^{N \times d} \quad (3)$$

For each image, we adopt a pre-trained convolutional neural network to extract image feature and also encode the result into d-dimensional space.

$$H^m = ReLU(w * CNN(Image) + b) \in \mathbb{R}^d \quad (4)$$

3.3 Decomposition Network (D-Net)

We focus on contextual contrast of multimodal sarcastic tweets and design the decomposition network (D-Net) to represent the commonality and discrepancy of image and text in high-level spaces.

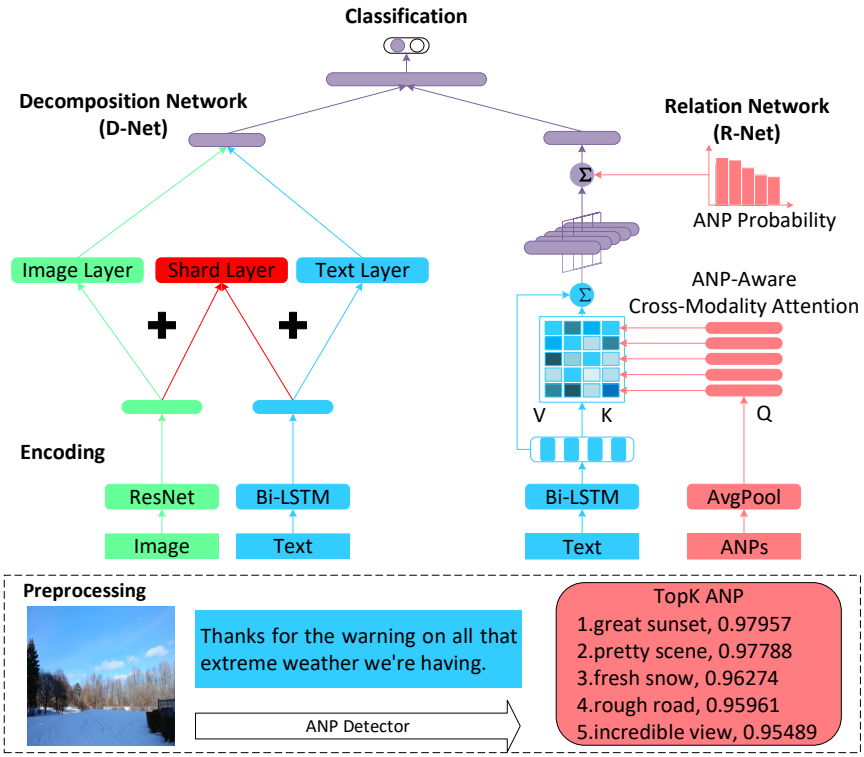

Figure 2: Overall architecture of our proposed D&R Net for multimodal sarcasm detection.

3.3.1 Cross-modality Decomposition

The D-Net breaks down the raw visual or textual representation into a shared subspace and unique visual or textual subspace through three layers. The shared layer tends to extract invariant shared features f^*_{shared} of image and text, and image or text layer is forced to decompose image or text into unique variant contrast features f^*_{unique}, which can be defined as

$$f^*_{shared} = W_{shared} f^* \in \mathbb{R}^{d_s} \quad (5)$$

$$f^*_{unique} = P^* f^* \in \mathbb{R}^{d_u} \quad (6)$$

where f^* is the feature of input modality $* \in \{image, text\}$, f^{image} is the raw image encoding representation H^m, f^{text} is the last hidden state h^w_T of BiLSTM which is used as the overall representation of text, and $W_{shared} \in \mathbb{R}^{d_s \times d}$, $P^* \in \mathbb{R}^{d_u \times d}$ are projection matrices of shared space, unique visual space and textual space.

3.3.2 Decomposition Fusion

In multimodal sarcastic tweets, we expect our model to focus more on the opposite between different modality information. Thus, we reinforce discrepancy between image and text, and on the contrary, weaken their commonality. Specifically, we combine the above unique variant contrast features as the cross-modality contrast representation.

$$r_{dec} = [f^{image}_{unique} \oplus f^{text}_{unique}] \in \mathbb{R}^{2d_u} \quad (7)$$

where $\oplus$ denotes the concatenation operation.

3.4 Relation Network (R-Net)

We propose the relation network (R-Net) to fully capture the contextual association between image and text from multiple views.

3.4.1 ANP-Aware Cross-Modality Attention

The relationship between image and text is usually multi-coupled, that is text may involve multiple entities in images, whereas different regions of the image may also involve different text words. We have already extracted multiple ANPs as the visual semantic information, which is beneficial to model multi-view associations between image and text according to different views of ANPs. Thus, we propose the ANP-aware cross-modality attention layer to align textual words and ANPs via utilizing each ANP to query each textual word and computing their pertinence.

We first calculate the cross interactive attention matrix $S \in \mathbb{R}^{N \times T}$ to measure how text words and image ANPs relate.

$$S = H^p W (H^w)^T \quad (8)$$

where $W \in \mathbb{R}^{d \times d}$ is the parameter of bi-linear function, and each score $s_{ij} \in S$ indicates the semantic similarity between i-th ANP encoding $h_i^p \in H^p$ and j-th text word encoding $h_j^w \in H^w$.

We then compute the cross-modality attention weight α_j^i of i-th ANP for j-th textual word by normalizing the i-th row of attention matrix S, and calculate the weighted average of textual hidden states as the i-th ANP-aware textual representation $r^i \in \mathbb{R}^d$:

$$\alpha_j^i = \frac{e^{s_{ij}}}{\sum_{j=1}^{T} e^{s_{ij}}} \tag{9}$$

$$r^i = \sum_{j=1}^{T} \alpha_j^i h_j^w \tag{10}$$

Thus, we query the text N times with different ANPs to get multi-view textual representations $[r^1, r^2, \ldots, r^N]$. Our proposed ANP-aware cross-modality attention mechanism is a variant of multi-head attention (Vaswani et al., 2017) and can be considered as the cross-modality adaptation of topic-aware mechanism (Wei et al., 2019), modeling the cross-modality association between image and text from multiple ANP-aware points. Next, we detail how to fuse such representations to get the final text representation.

3.4.2 ANP-Probability Fusion

We extract ANPs from each image and only select the Top N ANPs according to their extracted probability values $[p_1, p_2, \ldots, p_N]$. Hence, different textual representations should be influenced by different ANP probability values. Thus, we get the final cross-modality association representation $r_{rel} \in \mathbb{R}^d$ by calculating weighted average of these ANP-aware textual representations $[r^1, r^2, \ldots, r^N]$ according to the related normalized ANP probability distributions.

$$\beta^i = \frac{p_i}{\sum_{k=1}^{N} p_k} \tag{11}$$

$$r_{rel} = \sum_{i=1}^{N} \beta^i r^i \tag{12}$$

3.5 Sarcasm Classification

Finally, we feed the above acquired cross-modality contrast and semantic association representations, denoted as r_{dec} and r_{rel} respectively, into the top fully-connected layer and use the sigmod function for binary sarcasm classification.

$$\hat{y} = Sigmod\left(w_s[r_{dec} \oplus r_{rel}] + b_s\right) \tag{13}$$

where $w_s \in \mathbb{R}^{1 \times (2d_u + d)}$, $b_s \in \mathbb{R}^1$ are the parameters of fully-connected layer.

3.6 Optimization

Our model optimizes two losses, including classification loss and orthogonal loss.

We use cross entropy loss function as the sarcasm classification loss:

$$\mathcal{L}_c = -\sum_i y_i \log \hat{y}_i \tag{14}$$

where y_i is the ground truth of i-th sample (i.e., 1 for sarcasm and 0 for non-sarcasm), and $\hat{y}_i$ is the predicted label of our model.

In D-Net (Subsection 3.3), we share the same matrix for both image and text to ensure projecting them into the same subspace. Besides, in initialization and training process, to ensure that the decomposed unique subspaces are unrelated or in conflict with each other, we impose their projection matrices P^* with the additional orthogonal constraint for the shared projection matrix W_{shared}.

$$W_{shared}^T P^* = 0 \ (* \in \{image, text\}) \tag{15}$$

We convert these orthogonal constraints into the following orthogonal loss:

$$\mathcal{L}_o = \sum_{* \in \{image, text\}} \left\| W_{shared}^T P^* \right\|_F^2 \tag{16}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm.

We finally minimize the combined loss function:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_o \tag{17}$$

where λ is the weight of orthogonal loss.

4 Experiments

4.1 Dataset

We use a publicly available dataset constructed by Cai et al. (2019) to evaluate our model for multimodal sarcasm detection. Each sample in this dataset is image-text pair. This dataset is collected from Twitter by querying special hashtag (e.g. *#sarcasm, #sarcastic, #irony, #ironic* etc.) for positive samples (i.e. sarcasm) and the others without such hashtags as negative samples (i.e. non-sarcasm). The dataset has been divided into training set (80%), development set (10%) and test set (10%). Details are given in Table 2.

Method	Inputs	Evaluation Metric			
		F1	*P*	*R*	*Acc*
MLP+CNN (Schifanella et al., 2016)	1-grams + Image	75.83	79.52	72.47	81.61
Hierarchical FM (Cai et al., 2019)	Word2vec + Image + Attribute	80.18	76.57	84.15	83.44
D&R Net	Word2vec + Image + ANPs	**80.60**	77.97	83.42	**84.02**

Table 1: Comparative results with multimodal baselines

	Train	Dev	Test
Sarcasm	8642	959	959
Non-Sarcasm	11174	1451	1450
All	19816	2410	2409

Table 2: Statistics of the dataset

4.2 Implementation Details

For fair comparison, we adopt the same data preprocessing used in (Cai et al., 2019), replacing the mentions with a certain symbol $user$, cleaning up samples in which the regular words include 'sarcasm' related words (e.g. *sarcasm, sarcastic, irony, ironic*) and co-occur words (e.g. *jokes, humor, exgag*), and removing the stop words and URLs. We separate the text sentence by NLTK toolkit and embed each token into 200-dimensional word embedding by GloVe (Pennington et al., 2014). For image preprocessing, we first resize it into 224*224 and utilize pre-trained ResNet (He et al., 2016) to extract image feature. We also use SentiBank toolkit[1] to extract 1200 ANPs and select the Top 5 ANPs as the visual semantic information of each image. We encode the multimodal inputs into 200-dimensional hidden space, and set the dimension of invariant shared feature to 40, the dimension of unique variant contrast feature to 40, Finally, we optimize our model by Adam update rule with learning rate 0.01, mini-batch 128, and weight of orthogonal loss 0.5. The dropout and early-stopping tricks are utilized to avoid overfitting.

4.3 Comparison with multimodal baselines

Our work focus on the multimodal sarcasm detection using image and text modalities. Thus, we compare our model with the only two existing related models using the same modalities.

- **MLP+CNN** (Schifanella et al., 2016) concatenates multimodal features generated by textual MLP layer and visual CNN model for sarcasm classification, which is the first work on multimodal sarcasm detection.
- **Hierarchical FM** (Cai et al., 2019) takes text, image and image attributes as three modalities and fuses them with a multimodal hierarchical fusion model, which is the state-of-the-art method in multimodal sarcasm detection task.

We compare our model with multimodal baseline models with the F1-score and Accuracy metrics. Table 1 shows the comparative results. The MLP+CNN model simply takes the multimodal sarcasm detection as a general multimodal classification task via directly concatenating multimodal features for classification. Thus, it gets the worst performance. Hierarchical FM performs better than MLP+CNN by incorporating additional attributes that provide the visual semantic information and generating better feature representations via a hierarchical fusion framework. However, these multimodal baselines pay more attention to the fusion of multimodal features. In contrast, our D&R Net captures the essence of multimodal sarcasm via modeling cross-modality contrast in the associated context and achieves the best performance.

4.4 Comparison with unimodal baselines

To further explore the effects of multimodal inputs for sarcasm detection, we compare our model with the representative text-based sarcasm detection models and an image-based baseline model.

- **ResNet** (He et al., 2016) is widely used in many image classification tasks with prominent performance. As there is no related work on image sarcasm detection, we fine-tune it for image sarcasm classification.
- **CNN** (Kim, 2014) is a well-known model for many text classification tasks, which captures n-gram features by multichannel parameterized sliding windows.
- **BiLSTM** (Graves and Schmidhuber, 2005) is a popular recurrent neural network to model text sequence and incorporate bidirectional context information.

[1] ee.columbia.edu/ln/dvmm/vso/download/sentibank.html

Method	Modality	Evaluation Metric			
		F1	*P*	*R*	*Acc*
ResNet	Image	65.13	54.41	70.80	64.76
CNN	Text	75.32	64.29	76.39	80.03
BiLSTM	Text	77.53	76.66	78.42	81.90
MIARN	Text	77.36	79.67	75.18	82.48
D&R Net	Image+Text	**80.60**	77.97	83.42	**84.02**

Table 3: Comparative results with unimodal baselines

- **MIARN** (Tay et al., 2018) learns the intra-sentence relationship and sequential composition of sarcastic text, which is state-of-the-art method for text-only sarcasm detection.

We use F1-score and Accuracy as the evaluation metrics. Table 3 shows the comparative results of our model and these unimodal baseline models. Though ResNet demonstrates the superior performance in many image classification tasks, it performs relatively poor in sarcasm detection task. It is because that the sarcasm intention or visual contrast context in the image is usually unobvious. CNN and BiLSTM just treat the sarcasm detection task as a text classification task, ignoring the contextual contrast information. Thus, their performances are worse than MIARN, which focuses on textual context to model the contrast information between individual words and phrases. However, due to the nature of short text, relying on textual information is often insufficient, especially in multimodal tweets where cross-modality context relies the most important role. Our D&R Net performs better than unimodal baselines, demonstrating the usefulness of modeling multiple modality information in providing additional cues through reasoning contextual contrast and association.

4.5 Ablation Study

To evaluate the performance of each component used in our D&R Net, we conduct the detailed ablation studies on various variants of our model. The ablation results are shown in Table 4.

In general, we find those variants underperform our model. The most obvious declines come from the direct removal of our two core modules, D-Net and R-Net (see row 1, 3). Comparing these two variants, we find that removing D-Net has greater performance drop than removing R-Net. This suggests that modeling the cross-modality contrast in D-Net is more useful than cross-modality association in R-Net. After removing the D-Net, the model only accepts the text and ANPs inputs. Thus we further incorporate image information via directly concatenating image encoding in the final fusion layer (see row 2). The improvement compared with - D-Net shows the effectiveness of using image modality for multimodal sarcasm detection. Similarly, we also add the representation of ANPs to the fusion layer after removing the R-Net module (see row 4). However, the performance unexpectedly continues to decrease. One possible reason for this is that the fusion of ANPs affects the original decomposition results in spite of using triple inputs. It is worth mentioning that replacing our ANPs with noun attributes used in (Cai et al., 2019) underperforms our model (see row 5). This result indicates that ANPs are more useful in modeling semantic association between image and text compared with noun attributes. It is because that the adjective-noun words in ANPs are more semantically informative than noun-only words. Finally, we notice that our ANP-probability fusion (i.e. ANP-P.F.) strategy provides a means for obtaining reasonable performance compared with standard pooling operations, MaxPool and AvgPool (see row 6, 7), with ANP-probability weighted average performing the best.

	Variant	Evaluation Metric			
		F1	*Acc*	△*F1*	△*Acc*
	D&R Net	**80.60**	**84.02**	-	-
1	- D-Net	77.63	82.27	-2.97	-1.75
2	+ ⊕Image	79.10	82.73	-1.50	-1.29
3	- R-Net	79.90	83.10	-0.70	-0.92
4	+ ⊕ANPs	78.68	83.11	-1.92	-0.91
5	- ANP, +Attribute	79.52	83.12	-1.08	-0.90
6	- ANP-P.F., +MaxPool	79.80	83.27	-0.80	-0.75
7	- ANP-P.F., +AvgPool	79.86	83.42	-0.74	-0.60

Table 4: Ablation results of our D&R Net

4.6 Case Study

In this section, we provide case studies through several practical examples to illustrate that our D&R Net really learns to reason multimodal sarcastic tweets with interpretability.

4.6.1 Illustrative Examples

Fig.3 shows some multimodal non-sarcasm and sarcasm examples that our model correctly predicts. For those text-only or image-only models, it's almost impossible to detect the sarcasm intention of Fig.3a and 3b. We also show the results of the extracted ANPs from each image and these ANPs actually provide useful information for sarcasm

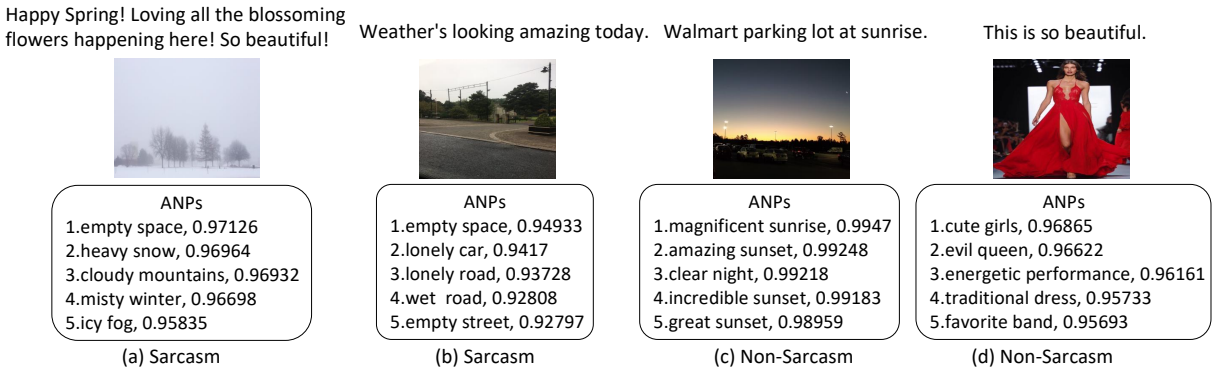

Figure 3: Examples of multimodal non-sarcasm and sarcasm tweets with extracted ANPs results.

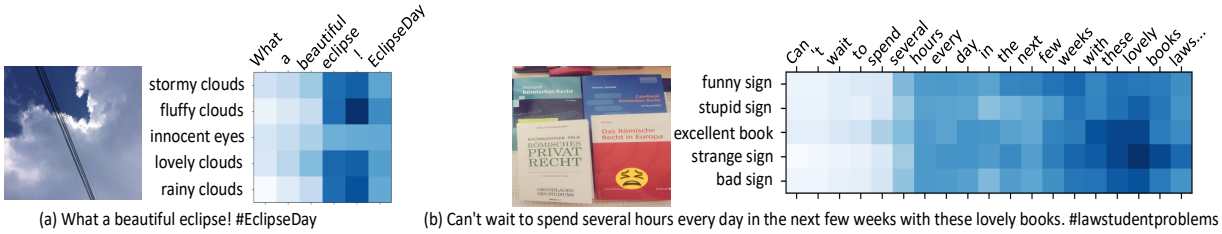

Figure 4: ANP-aware cross-modality attention visualization of multimodal sarcasm examples.

detection. For example, the ANPs *heavy snow*, *cloudy mountains*, *minsty winter* of Fig.3a show the great conflict with text word 'Spring', conveying the strong intention of sarcasm. In addition, our extracted ANPs are more semantically meaningful than the noun-only attributes used in (Cai et al., 2019). The *wet road* and *empty street* are more informative than noun-only words *road* and *street* in Fig.3b. The *cute girls* and *energetic performance* are more in line with the text words 'so beautiful' compared with noun-only words *girls* and *performance* in Fig.3d to discriminate between sarcasm and non-sarcasm.

4.6.2 Attention Visualization

Our proposed ANP-aware cross-modality attention mechanism explicitly calculates the cross interactive attention between text words and image ANPs, providing the explainable reasoning evidence for sarcasm detection. We further illustrate this attention mechanism by visualizing its outputs on two multimodal sarcastic tweets in Fig.4. The results show that our proposed attention mechanism works well for multimodal sarcasm detection by explicitly identify the relationship between image regions and text words. For instance, in Fig.4a, the user satirically mentions eclipse for too many clouds covering the sun. Our D&R Net accurately detects sarcasm intention via focusing on the text words 'eclipse', '!', 'EclipseDay' with multiple visual semantic ANP views: *stormy*, *fluffy*, *lovely* and *rainy clouds*. In Fig.4b, our model pays more attention to the textual phrase 'these lovely books' with *stupid sign, strange sign*, and *bad sign* ANPs which refer to the emoji in the attached image. Consequently, it is easy for our model to detect the sarcasm intention that the books are NOT 'lovely' at all.

5 Conclusion

In this paper, we identify the essential research issue in multimodal sarcasm detection. To model the cross-modality contrast in the associated context of multimodal sarcastic tweets, we propose the D&R Net to represent the commonality and discrepancy between image and text and multi-view semantic associations in cross-modality context. Our model is capable of reasoning multimodal sarcastic tweets with word-level interpretation. Experimental results on a public dataset show that our model achieves the state-of-the-art performance compared with the existing models.

Acknowledgments

This work is supported in part by the Ministry of Science and Technology of China under Grants #2016QY02D0305 and #2018ZX10201001, and National Natural Science Foundation of China under Grants #11832001, #71702181 and #71621002.

References

Silvio Amir, Byron C Wallace, Hao Lyu, Paula Carvalho, and Mario J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of CoNLL*, pages 167–177.

David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. In *Proceedings of ICWSM*, pages 574–577.

Christos Baziotis, Athanasiou Nikolaos, Pinelopi Papalampidi, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 3: Tracking ironic tweets using ensembles of word and character level attentive rnns. In *Proceedings of SemEval*, pages 613–621.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of ACL*, 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 ACL*, pages 4619–4629.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of CoNLL*, pages 107–116.

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. In *Proceedings of EMNLP*, pages 1615–1625.

Aniruddha Ghosh and Tony Veale. 2017. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *Proceedings of EMNLP*, pages 482–491.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of ACL*, pages 581–586.

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 *Proceedings of CVPR*, pages 770–778.

Irazú Hernández-Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying basic features from sentiment analysis for automatic irony detection. In *Proceedings of IbPRIA*, pages 337–344.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of ACL-IJCNLP*, pages 757–762.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *Proceedings of EMNLP*, pages 1006–1011.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP*, pages 1746–1751.

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

Lotem Peled and Roi Reichart. 2017. Sarcasm sign: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of ACL*, pages 1690–1700.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

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

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of WSDM*, pages 97–106.

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 EMNLP*, pages 704–714.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of ACM MM*, 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 ACL*, pages 1010–1020.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of ICWSM*, pages 162–169.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Penghui Wei, Wenji Mao, and Chen Guandan. 2019. A topic-aware reinforced model for weakly supervised stance detection. In *Proceedings of AAAI*, pages 7249–7256.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING*, pages 2449–2460.