

Utterance-Level Incongruity Learning Network for Multimodal Sarcasm Detection

Liujing Song^{*†}, Zefang Zhao^{*†}, Yuxiang Ma[§], Yuyang Liu[¶] and Jun Li^{*‡}

^{*}Computer Network Information Center, Chinese Academy of Sciences, Beijing, China

[†]University of Chinese Academy of Sciences, Beijing, China

[§]School of Computer and Information Engineering, Henan University, Kaifeng, China

[¶]Institute of Medical Information, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
{songliujing, zhaozefang, lijun}@cnic.cn, y.x.ma@hotmail.com, liu.yuyang@imicams.ac.cn, [‡]Corresponding Author

Abstract—With the exponential growth of user-generated online videos, multimodal sarcasm detection has recently attracted widespread attention. Despite making significant progress, there are still two main challenges: 1) previous works primarily relied on word-level feature interactions to establish relationships between inter-modality and intra-modality, which could potentially lead to the loss of fundamental emotional information. 2) they obtained the incongruity information only interacted with textual modality, which may lead to the neglect of incongruities. To address these challenges, we propose a novel utterance-level incongruity learning network (ULIL) for multimodal sarcasm detection, where the multimodal utterance-level attention (M-ULA) and incongruity learning network (ILN) are the two core modules. First, we present M-ULA to interact with utterance-level multimodal information, complementing word-level features. Furthermore, ILN selects primary modality and auxiliary modality automatically, and leverages cross-attention and self-attention to learning incongruity representations. We conduct extensive experiments on public datasets, and the results indicate that our proposed model achieves state-of-the-art performance in multimodal sarcasm detection.

Keywords—multimodal sarcasm detection, utterance-level attention, incongruity learning

I. INTRODUCTION

Sarcasm is a sophisticated form of linguistic expression where the intended meaning often contrasts with the literal interpretation [1]. This implicit emotional expression poses significant challenges for natural language processing tasks such as sentiment analysis and opinion mining. Previous sarcasm detection methods have primarily focused on text modality for classification task [2], [3]. However, the proliferation of user-generated online videos presents both opportunities and challenges for multimodal sarcasm detection.

Recently, some works have conducted multimodal sarcasm detection by capturing incongruity between modalities, including textual, visual and audio. Castro et al. [4] first extracted each modality feature and then fused them through early fusion. Chauhan et al. [5] combined the inter-segment inter-modal attention representation and intra-segment inter-modal attention representation for multitasking. Wu et al. [6] constructed a scare mechanism to acquire incongruity information based on word-level. Despite these advances, multimodal sen-

timent feature extraction remains a challenging problem. First, relying solely on word-level features is insufficient to capture the complex interplay between modalities. Second, incongruity information may occur on arbitrarily inter-modality and intra-modality. Thereby, just exploring incongruity information from textual with other modalities may cause a loss of comprehensive inconsistent information.

To address the issues mentioned above, we present a Utterance-Level Incongruity Learning Network (ULIL) for multimodal sarcasm detection, which comprises two core modules, namely multimodal utterance-level attention (M-ULA) and incongruity learning network (ILN). M-ULA is a novel multimodal information interaction and feature fusion module that can thoroughly explore complex inter-modal and intra-modal relationships. ILN is a novel incongruity representation network that performs comprehensive multimodal sarcasm detection. Specifically, ILN first dynamically determines the primary modality based on its contribution to the detection task and then hierarchically fuses the auxiliary modalities via cross-attention and self-attention mechanisms to efficiently obtain the most useful incongruity information.

In summary, the main contributions of our work are summarized as follows:

- We propose a novel utterance-level incongruity learning network for multimodal sarcasm detection. The core module is multimodal utterance-level attention (M-ULA) which relies on utterance-level to extract unimodal and multimodal features and effectively explore the complex interrelationships present in multimodal data.
- To capture the incongruity information, we design an incongruity learning network (ILN) which consists of primary modality choosing, cross attention mechanism and self-attention mechanism to explore sufficient incongruity information.
- We conduct extensive experiments and comparisons on public dataset to demonstrate the superiority of the proposed method. Furthermore, we perform ablation experiments to demonstrate the rationality and necessity of the proposed approach.

Organization The rest of the paper is organized as follows. A general exploration of textual modality and multimodal sarcasm detection methods are given in Section II. Section III describes the ULIL framework. The fundamental experimental setup is depicted in IV. Following, the extensive results and analysis are show in V. Ablation studies are discussed in Section VI. Finally, Section VII concludes the proposed model.

II. RELATED WORK

A. Textual Modality Sarcasm Detection

Sarcasm detection in the textual modality has long been a research focal point in natural language processing. Early research relied on rule-based methods [7], [8] and feature engineering-based machine learning methods [9], [10] to represent incongruous emotional expressions. However, with the evolution of neural networks, deep learning methods like CNN [11], RNN [12], and GNN [3] have found widespread application in textual sarcasm detection. In recent years, Tay et al. [2] explored an attention-based neural model which attempts to acquire contrast and incongruity information. Xiong et al. [13] proposed a self-matching network to capture sentence incongruity by discovering word-to-word interactions. Lou and Liang et al. [3] first leveraged Graph Convolutional Network to draw a long-range incongruity pattern by modeling the affective and dependency features information. Nonetheless, manual graph construction can potentially lead to some errors. To alleviate this limitation, Wang et al. [14] adopted an iterative augmenting affective and dependency graph architecture to learn the incongruity graph structure interactively.

B. Multimodal Sarcasm Detection

With the rapid growth of online media, data modalities have become increasingly diverse, leading researchers to incorporate vision and audio modalities in addition to text modalities for multimodal sarcasm recognition. As for multimodal sarcasm detection involving textual and visual modalities, Cai et al. [1] created a public dataset and proposed a fusion model by fusing text features, image features and image attributes to address the task. Liang et al. [15] constructed a crossmodal graph for each text and region to draw the sentiment relations between textual and visual modalities directly. Qiao et al. [16] explored a local and global incongruity learning network to seek underlying consistency between textual and visual modalities. As for multimodal sarcasm detection involving textual, visual and audio tri-modalities, Castro et al. [4] argued that incorporating multimodal cues can enhance automatic sarcasm detection. Thereby they constructed a trimodal public dataset. Chauhan et al. [5] concatenated the inter-segment inter-modal attention representation and intra-segment inter-modal attention representation for multi-tasking. Wu et al. [6] adopted incongruity aware attention network which focuses on word-level to detect sarcasm by score mechanism.

III. PROPOSED METHOD

A. Problem Statement

Multimodal sarcasm detection uses valid information from different modalities to judge potential inconsistencies. In this research, three modalities sequences including text $I_t \in \mathbb{R}^{L_t}$, acoustic $I_a \in \mathbb{R}^{L_a}$ and visual $I_v \in \mathbb{R}^{L_v}$ are used as source data. Furthermore, we consider multimodal sarcasm detection as a binary classification task, where three sequences are taken as input, and sarcasm is the final prediction.

B. Overall Architecture

As illustrated in Fig. 1, our proposed utterance-level incongruity learning network for multimodal sarcasm detection method consists of three layers. For the utterance-level feature extraction layer, to capture richer effective information from both intra-modality and inter-modality, we first use the unimodal encoder and utterance-level attention to obtain text, audio and vision features respectively. Then we leverage multimodal utterance-level attention (M-ULA) to acquire interacted fusion features. For the incongruity learning network layer, we first need to determine the primary modality, then employ a cross-attention mechanism to capture the inconsistency information, and finally employ self-attention mechanism to obtain the enhanced inconsistency information. For the prediction layer, the multimodal sarcasm label is predicted through a softmax classifier.

C. Utterance-Level Feature Extractions

a) *Unimodal Feature Extractions*: This module includes three main functions: token-level unimodal feature extraction, alignment and utterance-level feature representation. For the textual modality, we employ a pre-trained BERT-base model to encode the input textual tokens into word embedding. For audio and vision modalities, referring to previous work, we take advantage of Long Short-Term Memory(LSTM) to extract the initial features, which are denoted as:

$$\begin{aligned} I'_t &= \text{BERT}(I_t) \\ I'_a &= \text{LSTM}(I_a) \\ I'_v &= \text{LSTM}(I_v) \end{aligned} \quad (1)$$

where $I'_f \in \mathbb{R}^{L_f \times d_f}$, $f \in \{t, a, v\}$, L_f is the feature length and d_f represents the feature dimension.

Moreover, we deploy a 1-dimensional convolution layer and a scale layer to align the dimension of feature vectors space, which is denoted as:

$$X_f = \text{Scale}(\text{Conv}(I'_f)) \quad (2)$$

where $\text{Scale}(\theta) = \frac{\theta}{\sqrt{\|\theta\|_2}}$, $X_f \in \mathbb{R}^{L_f \times d_f}$. It is worth noting that we standardize the dimensions of feature spaces from different modalities through convolution layer and scale, that is $d_t = d_a = d_v$.

Token-level feature interaction may lead to missing vital sentiment information. Therefore, we employ utterance-level attention(ULA) to learn the distinctive properties of unimodal modality. Nonetheless, the unimodal ULA module can only

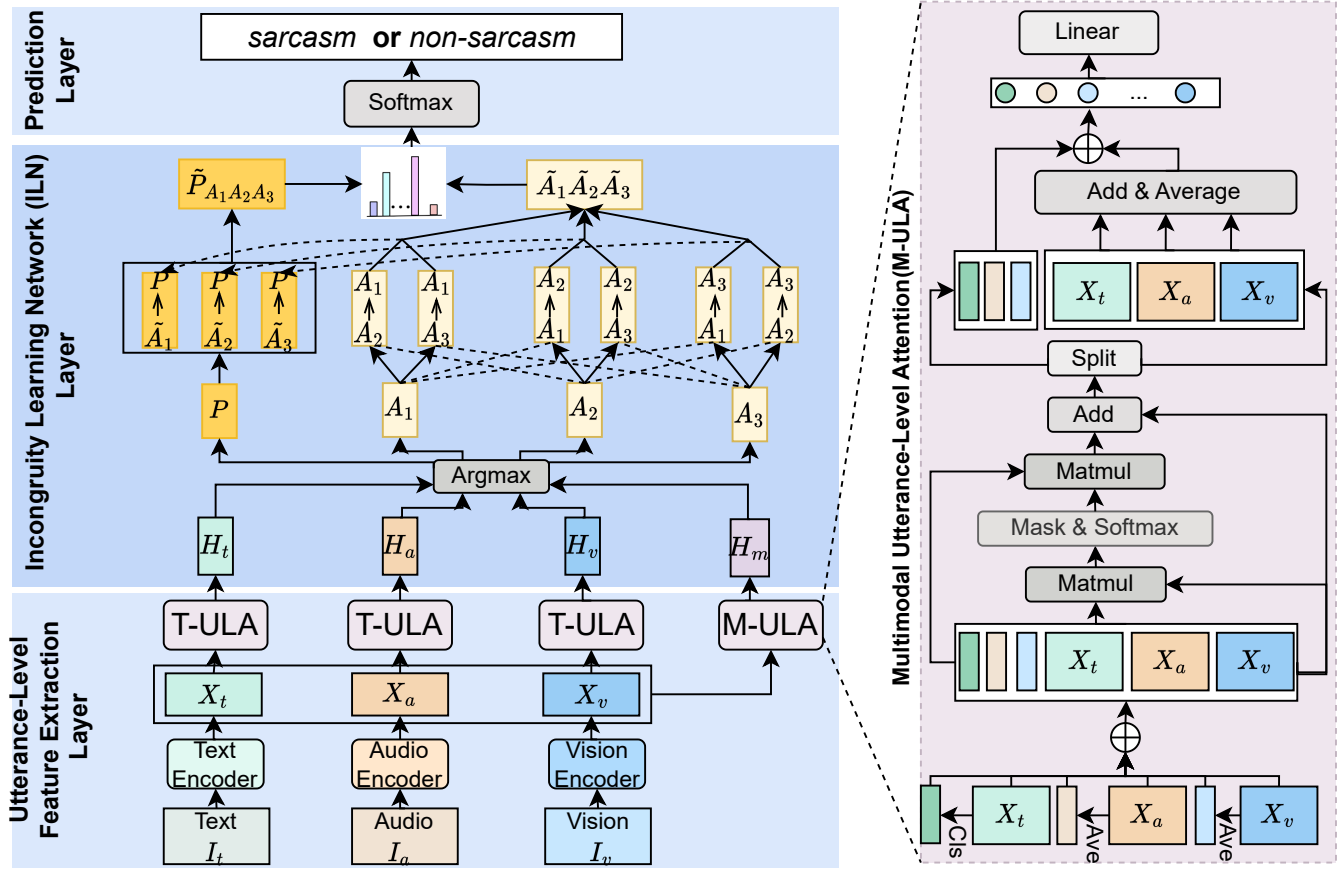


Fig. 1. The architecture of the proposed utterance-level incongruity learning network (ULIL) for multimodal sarcasm detection. At the bottom is the utterance-level feature extraction layer, which can extract unimodal features and interact with multimodal features by the Multimodal Utterance-Level Attention (M-ULA) module on the right. In the middle is the incongruity learning network layer, which can obtain incongruity information via cross-attention and self-attention.

explore internal relationships within the single modality. We extract unimodal features as input and obtain the utterance-level features representation is:

$$H_f = \text{ULA}(X_f) \quad (3)$$

where $H_f \in \mathbb{R}^d$. The unimodal ULA is a simplified version of the M-ULA. For detailed information, refer to the next section. Hence, we first acquire utterance-level features U_f through token-level features X_f . Subsequently, we can obtain the updated features H_f from each unimodal ULA.

$$H_f = \text{Linear}(|U'_f, \text{Ave}(X'_f)|) \quad (4)$$

b) Multimodal Feature Extraction: The main task of this module is to learn complex relationships of inter-modal, interact with multimodal information and gain the fusion feature representations with multimodal ULA. We feed X_t , X_a and X_v into the M-ULA and learn the fusion multimodal feature representation:

$$H_m = \text{ULA}(X_t, X_a, X_v) \quad (5)$$

The architecture of M-ULA is illustrated on the right of Fig.1. We design multimodal utterance-level attention (M-ULA) to explore complex intra-modal and inter-modal inter-

actions. We aim to learn a more enriched multimodal fused feature representation with abundant emotional information.

We take three token-level unimodal features (X_t , X_a , X_v) as input and gain their utterance-level features represent (U_t , U_a , U_v). We employ attention computation to establish fully connected correlations among these six emotion features and finally obtain the updated features represented by fusing. For the textual modality, we utilize the first vector of X_t as U_t , corresponding to the [CLS] token in BERT. For the audio and vision modalities, we can derive their initial utterance-level features by applying an average function. Next, we concatenate these six features to obtain the fused features:

$$X_m = |U_t, U_a, U_v, X_t, X_a, X_v| \quad (6)$$

where $X_m \in \mathbb{R}^{(3+L_t+L_a+L_v) \times d}$, and $|\cdot|$ denotes concatenation function.

During the training process, the utterance-level features of the three modalities interact with different feature information based on the attention scores ($att_{i,j}$), enabling their dynamic updates. We first learn the $query(Q) = W_q X_m$, $key(K) = W_k X_m$ and $value(V) = W_v X_m$, where $Q/K/V \in \mathbb{R}^{(3+L_t+L_a+L_v) \times d}$ and W_f are the weights,

$f \in \{t, a, v\}$. Therefore, the formula for the attention matrix of multimodal features is:

$$attention_{i,j} = Q \times K^\top \quad (7)$$

where $attention_{i,j} \in \mathbb{R}$ and $i, j \in [1, 3 + L_t + L_a + L_v]$.

Considering the impact of padding data on feature representation, we have devised a mask matrix M . It sets the positions of actual words and utterances to zero, while the positions of padding data are set to $-\infty$. The resulting attention score matrix is derived as follows:

$$att_score = \text{Softmax}(attention_{i,j} + M) \quad (8)$$

The updated multimodal feature representations $X'_m \in \mathbb{R}^{(3+L_t+L_a+L_v) \times d}$ are obtained using the residual function:

$$X'_m = att_score * V + X_m \quad (9)$$

Then, we split the interactive features into two sets, utterance-level features (U'_t, U'_a, U'_v) and token-level features (X'_t, X'_a, X'_v). We independently weight and sum the three learnable token-level features with weights W_t, W_a, W_v , obtaining the fused token-level feature values F_token . Next, by taking the average, we obtain the global token feature D_token , which is further fused with the utterance-level features to generate fusion features D_m . Eventually, after passing through several linear layers, we obtain the multimodal features representation H_m :

$$\begin{aligned} H_m &= \text{Linear}(D_m) \\ &= \text{Linear}(|U'_t, U'_a, U'_v, \text{Ave}(F_token)|) \\ &= \text{Linear}(|U'_t, U'_a, U'_v, D_token|) \end{aligned} \quad (10)$$

where $H_m \in \mathbb{R}^d$, $F_token = W_t X'_t + W_a X'_a + W_v X'_v$.

D. Incongruity Learning Network

a) *Primary Modality Selecting*: To assess the modality inconsistency, it is essential to first identify the primary modality. Referring to [17], we automatically assign weight values based on the varying contributions of each modality for sarcasm detection, with a higher contribution leading to a higher weight value. We update the weight values in each training batch while maintaining the sum of all trainable weight values equal to 1. We obtain the primary modality P and auxiliary modalities A_1, A_2, A_3 .

b) *Incongruity Learning by cross-attention*: In order to obtain incongruity from primary modality and auxiliary modalities, we leverage the cross-attention(CMA) mechanism to acquire interacted auxiliary modalities:

$$\begin{aligned} \tilde{A}_1 &= \text{CMA}(\tilde{A}_2 \rightarrow \tilde{A}_1) \oplus \text{CMA}(\tilde{A}_3 \rightarrow \tilde{A}_1) \\ \tilde{A}_2 &= \text{CMA}(\tilde{A}_1 \rightarrow \tilde{A}_2) \oplus \text{CMA}(\tilde{A}_3 \rightarrow \tilde{A}_2) \\ \tilde{A}_3 &= \text{CMA}(\tilde{A}_1 \rightarrow \tilde{A}_3) \oplus \text{CMA}(\tilde{A}_2 \rightarrow \tilde{A}_3) \end{aligned} \quad (11)$$

Subsequently, we can obtain the interacted primary modality representations:

$$\begin{aligned} \tilde{P}_{\tilde{A}_1} &= \text{CMA}(\tilde{A}_1 \rightarrow P) \\ \tilde{P}_{\tilde{A}_2} &= \text{CMA}(\tilde{A}_2 \rightarrow P) \\ \tilde{P}_{\tilde{A}_3} &= \text{CMA}(\tilde{A}_3 \rightarrow P) \end{aligned} \quad (12)$$

c) *Incongruity Learning by self-attention*: We concatenate interacted primary modality $\tilde{P}_{\tilde{A}_1}$, $\tilde{P}_{\tilde{A}_2}$ and $\tilde{P}_{\tilde{A}_3}$, and then we employ a self-attention(SA) mechanism to identify its salient components as the ultimate primary modality representation:

$$\tilde{P}_{\tilde{A}_1 \tilde{A}_2 \tilde{A}_3} = \text{SA}(\tilde{P}_{\tilde{A}_1} \oplus \tilde{P}_{\tilde{A}_2} \oplus \tilde{P}_{\tilde{A}_3}) \quad (13)$$

In the end, we obtain the multimodal inconsistency representation by concatenating auxiliary modalities $\tilde{A}_1, \tilde{A}_2, \tilde{A}_3$ and primary modality $\tilde{P}_{\tilde{A}_1 \tilde{A}_2 \tilde{A}_3}$:

$$Z = W_1 \tilde{A}_1 \oplus W_2 \tilde{A}_2 \oplus W_3 \tilde{A}_3 \oplus \tilde{P}_{\tilde{A}_1 \tilde{A}_2 \tilde{A}_3} \quad (14)$$

where W_1, W_2 and W_3 are the weight trainable parameters, which are learned by the modal itself to adjust the amount of auxiliary information to be extracted.

E. Multimodal Sarcasm Classification

The task of multimodal sarcasm detection is to predict the label $y \in \{\text{sarcasm, non-sarcasm}\}$. Therefore, the ultimate utterance-level incongruity representation is passed through a fully connected layer with a softmax activation function to produce a probability distribution y in the multimodal sarcasm decision space:

$$y = \text{softmax}(W_o Z + b_o) \quad (15)$$

where W_o and b_o are trainable parameters.

IV. EXPERIMENT

A. Dataset

TABLE I
DATASETS STATISTICS

	<i>Fr</i>	<i>BBT</i>	<i>TGG</i>	<i>Sa</i>	<i>total</i>
Sarcasm	152	140	39	14	345
Non-sarcasm	204	140	1	0	345
total	356	280	40	14	690

We evaluate our proposed model on the public multimodal sarcasm detection benchmark dataset¹ **MUSARD** published by [4]. The dataset comprises 690 utterances (345 sarcastic examples and 345 non-sarcastic examples) sourced from famous TV shows such as Friends (Fr), Big Bang Theory (BBT), The Golden Girls (TGG), and Sarcasmaholics (Sa). The detailed statistics of the experimental dataset are presented in Table I. Following [4], [6], We also evaluated our model in two experimental settings. One is the speaker-independent setup, where utterances collected from the Friends are used as testing data, and the remaining utterances are used for training data. The other is a speaker-dependent setup, where the dataset is split into five parts. In each of the five iterations, the i -th part is used as the testing, while the rest is used for training. As a result, we obtained five datasets.

¹<https://github.com/soujanyaoria/MUSARD>

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS FOR UNIMODAL AND MULTIMODAL SARCASM DETECTION

Modality	Method	Speaker-Dependent			Speaker-Independent		
		Precision(%)	Recall(%)	F1-score(%)	Precision(%)	Recall(%)	F1-score(%)
T	SMSD	61.6	61.0	61.1	51.7	48.2	47.0
	MIARN	64.7	64.0	63.9	60.4	55.2	54.0
	BERT	67.5	66.9	66.8	58.2	56.7	57.0
T,A,V	RAVEN	69.1	67.5	67.1	53.8	50.4	49.7
	LSTM(A)	67.3	66.7	66.3	56.7	54.1	54.0
	MFN	70.1	69.6	69.7	66.0	62.5	62.4
	EF-Concat	71.2	70.8	70.8	64.3	63.1	63.3
	IAIE	72.1	71.6	72.0	66.0	65.5	65.6
	IWAN	75.2	74.6	74.5	71.9	71.3	70.0
T,A,V	ULIL	76.9	76.1	76.5	73.1	72.6	72.8

B. Baselines

To comprehensively evaluate the performance of our proposed model (**ULIL**), we compared it with the baselines in the following modality including Text modality models and Trimodal models.

Text modality models (T):

- **SMSD** [13]: investigating a self-matching network and a low-rank bilinear pooling method to extract incongruity.
- **MIARN** [2]: adopting a multi-dimensional intra-attention recurrent network to capture textual incongruity.
- **BERT** [18]: is a pre-trained model and achieves state-of-art methods in many natural language processing tasks.

Trimodal models (T, A, V):

- **RAVEN** [19]: uses nonverbal sub-networks and gated modality-mixing network to shift word representations.
- **LSTM(A)** [20]: captures word-level multimodal features by the gated multimodal embedding LSTM with temporal attention.
- **MFN** [21]: considers and dynamically models interactions between modalities within a neural architecture over time.
- **EF-Concat** [4]: is an utterance-level multimodal sarcasm detection method by SVM classifier.
- **IAIE** [5]: exploring inter-segment inter-modal attention and intra-segment inter-modal attention to fuse the multimodal features.
- **IWAN** [6]: is an incongruity-aware attention network which focuses on word-level incongruity between three modalities by a scoring mechanism.

C. Settings

Evaluation Metrics: Following previous work [6], we report the weighted *precision*, *recall* and *F1-scores* to evaluate the performance of baselines and our model.

Implementation Details: Our proposed method is implemented by PyTorch [22] and all of the experiments in the paper are employed on the NVIDIA Tesla V100 GPU. The hyperparameters are the same for the two experimental configurations. The proposed ULIL uses the Adam optimizer. The learning rate is set to 0.001. The batch size is set to 32. To mitigate overfitting, we set the dropout rate to 0.1. We set the dimension of the feature space after convolution to 300.

V. RESULTS AND ANALYSIS

The comparative evaluation experiment results of our proposed model and all baselines are shown in Table II. We can draw the following conclusions: 1) Our proposed ULIL outperforms existing baselines on all metrics, confirming the effectiveness of our model in multimodal sarcasm detection. 2) All of the trimodal sarcasm detection approaches are superior to just text modality detection methods, demonstrating the complementarity of text, image, and audio modalities, thereby enhancing the performance of sarcasm detection. 3) All models exhibit lower performance in the speaker-independent setup compared to the speaker-dependent one. This is attributed to the absence of speaker overlap between the training and testing sets in the speaker-independent configuration, making it more challenging than the speaker-dependent setting. 4) In both the speaker-dependent and speaker-independent settings, our model ULIL achieves an F1-score value that is 2% and 2.8% higher than the most recent baseline IWAN, respectively.

In summary, the experimental results validate the superiority of the proposed ULIL model for multimodal sarcasm detection. The performance improvement can be attributed to the advancement of ULIL. First, in the feature extraction layer, it not only extracts unimodal features at the utterance level but also fuses multimodal features at the utterance level. Second, it utilizes cross-attention and self-attention to extract information about inconsistency between modalities, leading to more accurate sarcasm detection.

VI. ABLATION STUDIES

To further analyze our proposed model and assess the necessity of the three modalities, we conduct ablation experiments on the dataset with different modalities removed. As text modality is the most fundamental, we separately removed audio and vision modalities and recorded the experimental results in Table III. Removing either audio or vision modality leads to a performance drop, indicating the crucial importance of these two modality features for the sarcasm detection task. Audio and vision modality features complement text features, enabling a more comprehensive capture of inconsistent features.

On the other hand, to further analyze the effectiveness of the main modules in ULIL, we removed the primary modules,

TABLE III
ABLATION STUDIES ABOUT DIFFERENT MODALITIES

Modality	Speaker-Dependent			Speaker-Independent		
	P	R	F1	P	R	F1
T,V	69.3	69.2	69.2	63.5	63.8	63.6
T,A	71.1	70.9	71.0	60.6	61.2	60.9
T,A,V	76.9	76.1	76.5	73.1	72.6	72.8

TABLE IV
EXPERIMENTAL RESULTS ABOUT MAIN MODULES

Modality	Speaker-Dependent			Speaker-Independent		
	P	R	F1	P	R	F1
w/o $\mathcal{U}$	74.8	74.1	74.4	69.6	70.3	69.9
w/o $\mathcal{I}$	75.9	75.1	75.5	70.6	70.9	70.7
ULIL	76.9	76.1	76.5	73.1	72.6	72.8

the multimodal utterance-level attention module (w/o $\mathcal{U}$) and the incongruity representation module (w/o $\mathcal{I}$). According to the results in Table IV, it is evident that the performance deteriorates when the utterance-level attention module is not used, underscoring its capability to capture crucial sentiment information. By comparing w/o $\mathcal{I}$ with ULIL, we can deduce the significance of incongruity feature representation in multimodal sarcasm detection.

VII. CONCLUSION

In this paper, we investigate a comprehensive and robust approach that employs an Utterance-level Incongruity Learning Network (ULIL) to address multimodal sarcasm detection tasks. Unlike previous works, we extract utterance-level unimodal features and multimodal features fused by a multimodal utterance-level attention. Additionally, we explore an incongruity learning network for sarcasm detection. Specifically, considering the sarcasm utterances process incongruous characteristics, we deploy a cross-attention and self-attention mechanism to acquire incongruities between primary modality and auxiliary modalities. Our proposed model has been experimentally tested on a public benchmark dataset, demonstrating a significant performance improvement compared to a series of baselines for the multimodal sarcasm detection task.

VIII. ACKNOWLEDGMENTS

This work was partially supported by the National Key Research and Development Project of China (No. 2022YFF0903203) and the Key Scientific and Technological Project of Henan Province, China (No. 222102210040).

REFERENCES

- [1] Y. Cai, H. Cai, and 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.
- [2] Y. Tay, L. A. Tuan, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," *arXiv preprint arXiv:1805.02856*, 2018.
- [3] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, "Affective dependency graph for sarcasm detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1844–1849.
- [4] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," *arXiv preprint arXiv:1906.01815*, 2019.
- [5] D. S. Chauhan, S. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? a multi-task learning framework for multimodal sarcasm, sentiment and emotion analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4351–4360.
- [6] Y. Wu, Y. Zhao, X. Lu, B. Qin, Y. Wu, J. Sheng, and J. Li, "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE MultiMedia*, vol. 28, no. 2, pp. 86–95, 2021.
- [7] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and 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.
- [8] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 757–762.
- [9] D. I. H. Farias, V. Patti, and P. Rosso, "Irony detection in twitter: The role of affective content," *ACM Transactions on Internet Technology (TOIT)*, vol. 16, no. 3, pp. 1–24, 2016.
- [10] N. Pawar and S. Bhingarkar, "Machine learning based sarcasm detection on twitter data," in *2020 5th international conference on communication and electronics systems (ICCES)*. IEEE, 2020, pp. 957–961.
- [11] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: <https://aclanthology.org/D14-1181>
- [12] S. Hiai and K. Shimada, "Sarcasm detection using rnn with relation vector," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 15, no. 4, pp. 66–78, 2019.
- [13] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *The world wide web conference*, 2019, pp. 2115–2124.
- [14] X. Wang, Y. Dong, D. Jin, Y. Li, L. Wang, and J. Dang, "Augmenting affective dependency graph via iterative incongruity graph learning for sarcasm detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 4702–4710.
- [15] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and 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.
- [16] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie, "Mutual-enhanced incongruity learning network for multi-modal sarcasm detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9507–9515.
- [17] Y. Wang, Y. Li, P. Bell, and C. Lai, "Cross-attention is not enough: Incongruity-aware multimodal sentiment analysis and emotion recognition," *arXiv preprint arXiv:2305.13583*, 2023.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [19] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.
- [20] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 163–171.
- [21] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.



Liuqing Song received his Bachelor's degree from Zhengzhou University. Now she is a Ph.D. student in the University of Chinese Academy of Sciences. His research focuses on natural language processing and data mining.



Zefang Zhao received his Bachelor's degree from Taiyuan University of Technology. Now he is a Ph.D. student in the University of Chinese Academy of Sciences. His research focuses on deep learning, natural language process and sentiment analysis.



Yuxiang Ma is currently an associate professor in the School of Computer and Information Engineering at Henan University. He received the B.S. degree from Henan University in 2013, and the Ph.D. degree from the Computer Network Information Center, Chinese Academy of Sciences in 2019. His main research interests include network security, mobile computing, and privacy enhancement technologies.



Yuyang Liu received the Ph.D. degree from University of Chinese Academy of Sciences. He is currently a Research Associate Professor in Institute of Medical Information, Chinese Academy of Medical Sciences. His research interests include complex medical knowledge networks, clinical decision support system, and various data mining and artificial intelligence applications across medical informatics.



Jun Li is a research fellow and doctoral supervisor at the Computer Network Information Center of Chinese Academy of Sciences, specially appointed researcher of Chinese Academy of Sciences. His main research interests are artificial intelligence and big data technical applications and future Internet architecture.