Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection

Binghao Tang, Boda Lin, Haolong Yan, Si Li∗
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{tangbinghao,linboda,yanhaolong,lisi}@bupt.edu.cn

Abstract
Multimodal sarcasm detection aims to identify sarcasm in the given image-text pairs and has wide applications in the multimodal domains. Previous works primarily design complex network structures to fuse the image-text modality features for classification. However, such complicated structures may risk overfitting on in-domain data, reducing the performance in out-of-distribution (OOD) scenarios. Additionally, existing methods typically do not fully utilize cross-modal features, limiting their performance on in-domain datasets. Therefore, to build a more reliable multimodal sarcasm detection model, we propose a generative multimodal sarcasm model consisting of a designed instruction template and a demonstration retrieval module based on the large language model. Moreover, to assess the generalization of current methods, we introduce an OOD test set, RedEval. Experimental results demonstrate that our method is effective and achieves state-of-the-art (SOTA) performance on the in-domain MMSD2.0 and OOD RedEval datasets. The source code and RedEval are available at https://github.com/TangBinghao/naacl2024.

1 Introduction
Sarcasm is a linguistic phenomenon of verbal irony where the literal meaning contradicts the real intent of the speaker. Sarcasm detection aims to identify the actual sentiment of the user and can be widely applied in various scenarios such as public opinion mining (Pang et al., 2008; Riloff et al., 2013) and social media analysis (Tsur et al., 2010). Recently, due to the rapid surge of multimodal data on social media, multimodal sarcasm detection has gained increasing attraction and significance (Cai et al., 2019; Xu et al., 2020; Pan et al., 2020; Wang et al., 2020; Liang et al., 2021, 2022; Pramanick et al., 2022; Liu et al., 2022a; Tian et al., 2023; Qin et al., 2023). As shown in Figure 1, the given image-text pair is sarcastic because the image fails to show a good execution of “replacing grass” while the text describes it as a “wonderful job”.

Figure 1: Illustration of multimodal sarcasm detection. The left of the figure conveys a sarcastic meaning with the contrast between “grass replacing” and “wonderful job”. The right part displays an image of “beautiful autumn”, which is semantically consistent and shows no sarcasm. Previous works rely on complex model structures for feature fusion followed by classification, whereas our method generates answers based on LLMs.
cross-modal features and limit the performance in in-domain situations. However, effectively balancing and integrating cross-modal feature interactions remains a critical challenge in enhancing the generalization and robustness of multimodal sarcasm detection models.

Fortunately, significant progress has been made in various NLP generation tasks with the development of Large Language Models (LLMs) (Ouyang et al., 2022). Further leveraging LLMs and extending them to multimodal domains, the Multimodal Large Language Models (MLLMs) (Zhu et al., 2023; Chen et al., 2023; Liu et al., 2023a,b) have also significantly improved various multimodal tasks and show great generalization.

Therefore, to build a more reliable multimodal sarcasm detection model, we redefine multimodal sarcasm detection as a generative task to take advantage of the powerful MLLMs. To further leverage and enhance the performance of MLLMs, we design a detailed instruction template and propose a simple yet effective demonstration retrieval module. Furthermore, considering the lack of research assessing the generalization of current multimodal sarcasm detection models, we collect multimodal data from other social media platforms and propose a new test set named RedEval for OOD evaluation. In summary, our contributions are as follows:

• To the best of our knowledge, we are the first to attempt to explore the generalization of multimodal sarcasm models and the first to redefine multimodal sarcasm detection as a generative task utilizing MLLMs for better balance and integration of cross-modal interactions.

• We propose a new test dataset, RedEval, comprising image-text pairs collected from other social media, to assess the generalization of existing models, aiming to construct more reliable multimodal sarcasm models.

• We design an instruction template and a retrieval module to further enhance our generative multimodal sarcasm model. Experimental results on MMSD2.0 and RedEval demonstrate that our method is effective in both in-domain and OOD situations and achieves SOTA performance.

2 Related Work

2.1 Multimodal Sarcasm Detection

Traditional sarcasm detection task aims to identify the sentiments of users and detect the presence of sarcasm from textual modality (Zhang et al., 2016; Tay et al., 2018; Babanejad et al., 2020). Due to the surge of multimodal data in social media, multimodal sarcasm detection has gradually attracted much attention.

Schifanella et al. (2016) firstly explores the multimodal sarcasm detection task by concatenating the textual and visual embeddings. Cai et al. (2019) proposes a hierarchical fusion network and releases a multimodal public dataset, i.e., MMSD. Subsequent studies further model the commonalities and incongruity between visual and textual modalities by a decomposition and relation network (Xu et al., 2020), BERT-based (Kenton and Toutanova, 2019) models through modified attention mechanisms (Pan et al., 2020; Wang et al., 2020), graph neural networks (Liang et al., 2021, 2022) and optimal transport (Pramanick et al., 2022). And Liu et al. (2022a) proposes a hierarchical framework with external knowledge enhancement for multimodal sarcasm detection. Recently, Tian et al. (2023) applies a dynamic routing network to model the cross-modal incongruity. Furthermore, Qin et al. (2023) discovers that existing models may overly rely on spurious textual cues rather than cross-modal features. This leads to the introduction of a new benchmark, MMSD2.0, and the proposal of a novel framework based on the vision-language pre-trained model CLIP (Radford et al., 2021) to capture sarcasm cues from diverse perspectives. Compared with prior works, our method redefines multimodal sarcasm detection as a generative task.

2.2 Multimodal Large Language Models

Large Language Models (LLMs) have achieved widespread success in the field of NLP. From early-stage models like BERT (Kenton and Toutanova, 2019) and GPT-2 (Radford et al., 2019) to more recent GPT-3 (Brown et al., 2020), instruct-GPT (Ouyang et al., 2022), and various other open-source large-scale language models, such as LLaMA (Touvron et al., 2023a) and LLaMA2 (Touvron et al., 2023b), there has been substantial development in the field of NLP, particularly in the area of natural language understanding and generation.

In the research of multi-modality, how to apply those powerful LLMs to multimodal tasks
has also gradually gained significant attribution. Early research like Frozen (Tsimpoukelli et al., 2021), achieves impressive performance by training a visual encoder to encode the image input as a prefix in a frozen pre-trained language model. BLIP (Li et al., 2022) pre-trains a multimodal mixture of encoder-decoder model to further boost vision-language tasks. BLIP2 (Li et al., 2023) proposes a Q-former to efficiently align visual features to LLMs. Additionally, other studies such as MiniGPT4 (Zhu et al., 2023; Chen et al., 2023), LLaVA (Liu et al., 2023a,b) and Qwen-VL (Bai et al., 2023) employ an adapter like a linear layer or multi-layer perceptron to further align the image features extracted from visual encoders like ViT (Dosovitskiy et al., 2020). We apply the multimodal sarcasm detection task to the generative framework of multimodal large language models to address the problems of insufficient generalization and inadequate reliance on multimodal features.

### 3 Methodology

In this section, we present the overview of our method. We first present the brief task formulation and describe the MLLM-based generative framework. Then we detail our retrieval module and introduce the training and generation process.

#### 3.1 Task Formulation

Given image-text pairs \( \langle v_i, t_i \rangle \), where \( v_i \) is the \( i \)-th image input and \( t_i \) is the \( i \)-th text input. MLLM needs to generate the sarcasm label from the label set \( S = \{ \epsilon_p, \epsilon_n \} \) based on \( v_i \) and \( t_i \), where \( \epsilon_p \) and \( \epsilon_n \) are the positive and negative labels.

#### 3.2 Model Framework

For the generative MLLMs, we leverage LLaVA-1.5 (Liu et al., 2023a) as our backbone. LLaVA-1.5 adopts a multi-layer perceptron as the cross-modal projection to connect the vision encoder and large language model. Liu et al. (2023a) further pre-trains the vision-language connector on the 600K public image-text pairs instructions data, which shows the strong power of various multimodal tasks. As shown in Figure 2, given the image-text pair, we first retrieve the best demonstration from the training set. Then we obtain the visual features from both the demonstration and sample images using the visual encoder and adapter. For the demonstration text and sample text, we input them into LLM along with the visual features in the format of an instruction template.

#### 3.3 Retrieval Module

To better prompt MLLMs to generate the right answers, we introduce a retrieval module for MLLMs to search for demonstrations, aiming at further...
bridging the gap between MLLMs and the specific multimodal sarcasm detection task.

As shown in Figure 2, for the given image-text pairs \( \langle v_i, t_i \rangle \), we first obtain their corresponding embeddings by CLIP (Radford et al., 2021):

\[
\text{Emb}_v(i) = \text{CLIP}_{\text{vis}}(v_i) \quad (1)
\]
\[
\text{Emb}_t(i) = \text{CLIP}_{\text{text}}(t_i) \quad (2)
\]

where CLIP_{\text{vis}} and CLIP_{\text{text}} are the visual and textual encoder of CLIP.

For each sample \( \langle \text{Emb}_v(i), \text{Emb}_t(i) \rangle \), we calculate the cosine similarity of image and text modalities separately with the samples in the training set \( D_{\text{train}} \):

\[
\text{Sim}_v(i) = \frac{\text{Emb}_v(i) \cdot V}{|\text{Emb}_v(i)||V|} \quad (3)
\]
\[
\text{Sim}_t(i) = \frac{\text{Emb}_t(i) \cdot T}{|\text{Emb}_t(i)||T|} \quad (4)
\]

where \( V \) and \( T \) are the image and text embeddings from the training set \( D_{\text{train}} \).

Finally, we select the sample with the highest average similarity score as the corresponding demonstration:

\[
\text{Demon}(i) = \arg \max \frac{\text{Sim}_v(i) + \text{Sim}_t(i)}{2} \quad (5)
\]

where \( \text{Demon}(i) \) is the similarity score of demonstration of \( i \)-th image-text sample. Then the retrieved demonstration and the sample, after being processed through the instruction template, are input into LLM together, as shown in Figure 2.

### 3.4 Optimization Objective

Consistent with the loss calculation in autoregressive LLMs, we only compute the cross-entropy loss for the response of MLLM, i.e., the label of the image-text pair:

\[
\mathcal{L} = \sum_{i=1}^{n} - \log p_\theta(\epsilon_i|\text{instruction}_i) \quad (6)
\]

where \( \text{instruction}_i \) is the \( i \)-th instruction containing the information of image-text pair, \( \epsilon_i \) is the corresponding predicted label and \( \theta \) represents the parameters of the MLLM.

### 3.5 Constrained Decoding

For the generative MLLMs, the results may not fully comply with the requirements even when the output format is specified in the input instruction. This poses a challenge for the classification results of the multimodal sarcasm detection task. To address this issue, we implement constrained decoding (De Cao et al., 2020), ensuring that the model can only generate outputs based on the label set.

## 4 Experiments

### 4.1 Dataset

#### In-Domain Dataset

We evaluate our method on MMSD (Cai et al., 2019) and MMSD2.0 (Qin et al., 2023) datasets. The MMSD dataset originates from the image-text pairs collected by Cai et al. (2019) on Twitter\(^1\) and is randomly divided into training, validation, and test sets in the ratio of 80%, 10%, and 10% respectively. The MMSD2.0 (Qin et al., 2023) dataset is built upon MMSD, involving the removal of spurious cues and re-annotating unreasonable samples on the textual content. The statistics of the MMSD and MMSD2.0 datasets are shown in Table 1. For a fair comparison, we conduct the same data preprocessing on the MMSD dataset following previous works.

#### Out-of-Domain Dataset

To assess the generalization of current multimodal sarcasm models, we propose a new test dataset called RedEval. Considering that the existing image-text pairs in the MMSD and MMSD2.0 datasets are all from the same social media Twitter, we select image-text pairs from another social media platform Reddit\(^2\) as the out-of-distribution data. Specifically, we select image-text data from the “sarcasm” subreddit as positive sarcasm samples, and a certain number of samples from other subreddits such as “aww”, “funny”, “pics”, and “popular” as the non-sarcastic samples. Following Qin et al. (2023), we remove the emotions from the data. We also employ 3 graduate students to ensure the quality of the image-text pairs in RedEval aligns with the intended meaning of sarcastic labels. In detail, each image-text pair is reviewed by two annotators. They are required to predict the sarcasm label. Only samples that receive consistent predictions from two annotators are left. Moreover, we let another annotator predict the sarcasm labels and then calculated the Kappa coefficient with the gold labels. We obtained a Kappa value of 0.793, indicating a high level of consistency. The statistics of RedEval are shown in Table 2, the maximum and average lengths are different from MMSD and MMSD2.0.

---

1https://twitter.com/
2https://www.reddit.com/
<table>
<thead>
<tr>
<th>Datasets</th>
<th>Split</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
<th>Max Len</th>
<th>Min Len</th>
<th>Avg Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSD</td>
<td>Training</td>
<td>8,642</td>
<td>11,174</td>
<td>19,816</td>
<td>70</td>
<td>1</td>
<td>15.71</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>959</td>
<td>1,451</td>
<td>2,410</td>
<td>55</td>
<td>1</td>
<td>15.72</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>959</td>
<td>1,450</td>
<td>2,409</td>
<td>64</td>
<td>1</td>
<td>15.89</td>
</tr>
<tr>
<td>MMSD2.0</td>
<td>Training</td>
<td>9,576</td>
<td>10,240</td>
<td>19,816</td>
<td>66</td>
<td>1</td>
<td>13.42</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>1,042</td>
<td>1,368</td>
<td>2,410</td>
<td>55</td>
<td>4</td>
<td>13.64</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>1,037</td>
<td>1,372</td>
<td>2,409</td>
<td>52</td>
<td>4</td>
<td>13.52</td>
</tr>
</tbody>
</table>

Table 1: The statistics of MMSD and MMSD2.0. Len denotes the number of words in the corresponding dataset.

<table>
<thead>
<tr>
<th></th>
<th>Pos.</th>
<th>Neg.</th>
<th>Total</th>
<th>Max</th>
<th>Min</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>RedEval</td>
<td>395</td>
<td>609</td>
<td>1,004</td>
<td>54</td>
<td>1</td>
<td>7.35</td>
</tr>
</tbody>
</table>

Table 2: The statistics of RedEval. Pos. and Neg. are the positive and negative samples. Max, Min, and Avg are the number of words as mentioned in Table 1.

### 4.2 Experimental Settings

Based on LLaVA-1.5-7B (Liu et al., 2023a), we use "CLIP-ViT-L-336px" as the vision encoder and "Vicuna-v1.5-7B" as the LLM. We use the same vision encoder to obtain image and text embeddings in the retrieval module. We utilize "BLIP2-FlanT5-XL" to obtain image captions. Given the limitations of task-specific data and computational resources, we choose Parameter-Efficient-Fine-Tuning (PEFT) for the training stage. Specifically, we adopt LoRA (Hu et al., 2021) and inject the low-rank matrices as adapters into MLLM. The rank of the update matrices is 128 and the scaling factor of LoRA is 256. We freeze the vision encoder and fine-tune the vision-language connector and LLM following Liu et al. (2023a). The learning rate for the vision-language connector is 2e-5 and the learning rate for LLM is 2e-4. The batch size is 12 and the training epoch is 5. We adopt the constrained greedy search for inference. All models are trained on 2 NVIDIA 3090Ti GPUs and tested on a single NVIDIA 3090Ti GPU.

### 4.3 Baselines

Following prior works, we compare our method with unimodal and multimodal baselines for multimodal sarcasm detection. (4) RoBERTa (Liu et al., 2019) is a robustly optimized BERT (Kenton and Toutanova, 2019) pre-trained language model. (5) ChatGLM2-6B (Du et al., 2022) is an open bilingual language model based on the general language model framework, with 6.2 billion parameters. (6) LLaMA2-7B (Touvron et al., 2023b) is a foundation LLM pre-trained on 2 trillion tokens, with 7 billion parameters. We refer to ChatGLM2-6B and LLaMA2-7B as the LLM-based methods.

**Image-Modality Methods.** (1) ResNet (He et al., 2016) utilizes the image embedding that is produced by the pooling layer to detect sarcasm. (2) ViT (Dosovitskiy et al., 2020) is a pre-trained vision transformer model.

**Multi-Modality Methods.** (1) HFM (Cai et al., 2019) is a hierarchical network with multimodal fusion. (2) D&R Net (Xu et al., 2020) proposes a decomposition and relation network to model the relationship between image and text. (3) Att-BERT (Pan et al., 2020) adopts self-attention and co-attention mechanisms to model the intra-modality and inter-modality incongruity respectively. (4) InCrossMGs (Liang et al., 2021) utilizes in-modal and cross-modal graphs to capture sarcastic relations between two modalities. (5) CMGCN (Liang et al., 2022) proposes a fine-grained cross-modal graph architecture to capture sarcastic clues. (6) HKE (Liu et al., 2022a) uses a hierarchical graph-based framework and incorporates external knowledge like image captions for multimodal sarcasm detection. (7) DynRTNet (Tian et al., 2023) proposes a dynamic routing transformer network to capture the sarcastic clues from images and texts. (8) Multi-view CLIP (Qin et al., 2023) utilizes a framework based on CLIP (Radford et al., 2021) from image view, text view, and image-text interactions view for multimodal sarcasm detection, which is current State-Of-The-Art (SOTA) multimodal sarcasm model. (9) LLaVA1.5-7B (Liu et al., 2023a) adopts a multi-
Table 3: Experimental results on MMSD and MMSD2.0. The best results are highlighted in bold, and the second-best results are underlined. * denotes the experimental results from Qin et al. (2023). † means our method outperforms Multi-view CLIP significantly with \( p < 0.001 \). Compared with MMSD2.0, the performance of text-modality methods reaches the highest on MMSD, indicating that MMSD is not sufficient to measure the effectiveness of multimodal methods.

4.4 Main Results

Following Qin et al. (2023), we adopt accuracy (Acc.), macro-average precision (P), macro-average recall (R), and macro-average F1 score (F1) as metrics to assess the performance of our model. 

Datasets Discussion. As shown in Table 3, for the MMSD dataset, the performance of LLMs like ChatGLM2-6B and LLaMA2-7B in the text modality methods reaches a relatively high level, even outperforming the multimodal methods. It is consistent with the experimental result of RoBERTa reported in Qin et al. (2023). This suggests that there indeed exists a problem with the text modality data in MMSD, which undermines the dependency of multimodal methods on image-text modality features. Furthermore, we observe that the performance of ChatGLM2-6B, LLaMA2-7B, and LLaVA1.5-7B on MMSD is very similar. This suggests that the performance of models on MMSD may have already reached the upper limit, making further improvements challenging. As for the MMSD2.0 dataset, multimodal model approaches generally outperform unimodal methods. Also, LLM-based methods utilizing image captions as visual information inputs could achieve better performance. This indicates that MMSD2.0 strengthens the dependency on cross-modal features, preventing models from relying solely on textual infor-
<table>
<thead>
<tr>
<th>Model</th>
<th>MMSD for OOD</th>
<th>MMSD2.0 for OOD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%) P (%) R (%) F1 (%)</td>
<td>Acc. (%) P (%) R (%) F1 (%)</td>
</tr>
<tr>
<td><strong>Text-Only Methods</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatGLM2-6B (Du et al., 2022)</td>
<td>58.47 46.51 49.13 41.12</td>
<td>77.19 76.74 74.57 75.22</td>
</tr>
<tr>
<td>ChatGLM2-6B (Du et al., 2022)†</td>
<td>58.57 46.79 49.21 41.17</td>
<td>79.28 78.92 80.25 78.95</td>
</tr>
<tr>
<td>LLaMA2-7B (Touvron et al., 2023b)</td>
<td>58.57 46.79 49.21 41.17</td>
<td>79.48 78.66 79.66 78.93</td>
</tr>
<tr>
<td>LLaMA2-7B (Touvron et al., 2023b)†</td>
<td>58.67 47.35 49.34 41.40</td>
<td>81.38 80.47 80.60 80.53</td>
</tr>
<tr>
<td><strong>Multi-Modal Methods</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DynRT-Net (Tian et al., 2023)</td>
<td>58.57 47.06 49.25 41.35</td>
<td>74.80 75.58 76.69 74.66</td>
</tr>
<tr>
<td>Multi-view CLIP (Qin et al., 2023)</td>
<td>76.29 75.67 73.70 74.30</td>
<td>80.98 80.85 82.62 80.73</td>
</tr>
<tr>
<td>LLaVA-1.5-7B</td>
<td>61.25 52.63 57.68 47.13</td>
<td>82.77 83.66 82.25 82.44</td>
</tr>
<tr>
<td>Ours</td>
<td>59.16 49.70 48.67 41.47</td>
<td><strong>83.47†</strong> 83.12 82.60 82.83†</td>
</tr>
</tbody>
</table>

Table 4: Experimental results on RedEval. The best results are highlighted in bold, and the second-best results are underlined. The models are trained on the training set of MMSD and MMSD2.0 and tested on RedEval. † and ‡ denote as the same in Table 3.

4.5 Out-of-Domain Results

As shown in Table 4, models trained on the MMSD dataset perform poorly on the OOD dataset. Only Multi-view CLIP (Qin et al., 2023) shows relatively better performance, yet still experiences a decline of over 12 points of accuracy compared to the in-domain situation. This also indicates that MMSD indeed causes models to focus more on domain-specific data features, or even solely on textual modality features, significantly damaging the generalization of models.

Compared to MMSD, models trained on the MMSD2.0 dataset exhibit better cross-modality dependence and generalization, as shown in Table 4, which is consistent with Section 4.4. But compared to the in-domain situations, there is still a noticeable performance decline. It has been observed that the models that perform well on MMSD2.0 also tend to show decent performance on RedEval. However, models that perform well on MMSD but poorly on MMSD2.0 like DynRT-Net (Tian et al., 2023) exhibit poor performance in the OOD situation. For the OOD decline performance, we Compared to previous methods, both LLM-based methods like LLaMA2-7B and MLLM-based method LLaVA show excellent performance, even surpassing the previous SOTA methods like Multi-view CLIP. This suggests that LLMs and MLLMs indeed have better generalization capabilities compared to other models. Moreover, our method still achieves the best accuracy and F1 scores on the OOD dataset, which further demonstrates the effectiveness of our approach under the OOD condition.
Table 5: The performance of retrieved demonstration quality and LLM-based methods on the MMSD2.0 dataset. The best results are highlighted in bold. RM denotes the retrieval module in Figure 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>MMSD2.0</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>P (%)</td>
<td>R (%)</td>
<td>F1 (%)</td>
</tr>
<tr>
<td>Demon Quality</td>
<td>79.20</td>
<td>79.36</td>
<td>78.90</td>
<td>78.90</td>
</tr>
<tr>
<td>ChatGLM2-6B</td>
<td>80.08</td>
<td>80.52</td>
<td>81.04</td>
<td>80.04</td>
</tr>
<tr>
<td>ChatGLM2-6B w/RM</td>
<td>82.94</td>
<td>82.62</td>
<td>83.12</td>
<td>82.76</td>
</tr>
<tr>
<td>LLaMA2-7B</td>
<td>84.68</td>
<td>84.40</td>
<td>84.94</td>
<td>84.53</td>
</tr>
<tr>
<td>LLaMA2-7B w/RM</td>
<td>85.97</td>
<td>85.64</td>
<td>85.85</td>
<td>85.74</td>
</tr>
</tbody>
</table>

For the observed performance degradation on RedEval, we speculate that it may be due to the source difference of multimodal samples between the RedEval dataset and the in-domain datasets. As described in Section 4.1, the datasets come from different social media platforms at different times, resulting in inherent differences in data distribution. From the statistics in Table 2, the average word count of RedEval is also lower than that of the in-domain dataset, indirectly reflecting the discrepancy between the datasets. Such differences in datasets leading to performance discrepancies further underscore the necessity of validating the generalization of models under OOD scenarios.

5 Analysis

5.1 The Effectiveness of Retrieval Module

In this section, we further analyze the effectiveness of our proposed retrieval module. Despite the results of the MLLM-based method in the main results Table 3, we further analyze the demonstration quality and the effectiveness of the retrieval module on LLM-based methods. Given that the performance of LLM-based methods on the MMSD dataset may reach the performance ceiling, we conduct experiments only on the MMSD2.0 dataset. We retrieve demonstrations for the given image-text pairs, requiring the LLMs to predict the corresponding label of the current sample based on the given demonstration. We keep adopting the format of image captions for the image information inputs.

We first evaluate the quality of our retrieved demonstrations. Generally, when the given image-text pair and its corresponding demonstration belong to the same category, the performance of models is better (Liu et al., 2022b). Therefore, we calculate the evaluation metrics by comparing the labels of the demonstrations with the corresponding image-text pairs. As shown in Table 5, with only the retrieval module, the performance of the retrieved demonstrations MMSD2.0 even surpasses most of the current multimodal methods, which are fine-tuned on the full training set. This indicates that our proposed retrieval module is highly effective in retrieving candidate demonstrations.

For the LLM-based methods, after incorporating our proposed retrieval module, they both achieve a further improvement, as shown in Table 5. Moreover, the performance of the LLaMA2-7B model even surpasses the previous SOTA method Multi-view CLIP (Qin et al., 2023). This indicates that even a purely textual LLM can achieve excellent performance in multimodal sarcasm detection by converting images into captions and applying proper prompting and instructions. This also reveals that LLMs not only have immense applicability in the field of natural language processing but also possess significant potential in the realm of multimodal tasks.

5.2 Low-Resource Scenario

Following Qin et al. (2023), we explore the effectiveness of our method and LLM-based methods in low-resource scenarios of MMSD2.0. Specifically, we compare the performance of our method, LLM-based methods with the retrieval module, and the previous SOTA method on MMSD2.0, Multi-view CLIP (Qin et al., 2023) in low-resource scenarios.

As shown in Figure 3, our MLLM-based method does not outperform Multi-view CLIP until the data proportion is above 40%. This may be because our method is based on a large scale of parameters of LLM and could not be trained sufficiently with the very limited data, leading to relatively poor perfor-
mance. With the continuous increase in data proportion in low-resource scenarios, we can observe that our method significantly outperforms Multi-view CLIP by a large margin after 40%. This indicates that with the continuous improvement in data scale, the performance based on the large language models can also be greatly enhanced.

For the LLM-based methods, we observe that LLaMA2-7B does not outperform Multi-view CLIP with a very limited amount of low-resource data like 50%. Beyond 50%, the performance of LLaMA2-7B gradually approaches and surpasses Multi-view CLIP. In contrast, ChatGLM2-6B consistently performs at a lower performance level, highlighting the importance of the choice of base models. The performance trends of LLM methods, as shown in Figure 3, are consistent with our method. This indicates that these LLM-based methods all have limitations with very low resources but their performance increases to a higher level once the data scale reaches a certain scale. This reveals that for LLMs, choosing an appropriate data scale is crucially important for performance.

6 Conclusion

In this paper, we focus on the problems of insufficient OOD generalization and inadequate utilization of cross-modal features in current multimodal sarcasm detection models. To build a more reliable model, we propose a generative multimodal sarcasm detection model consisting of an instruction template and a demonstration retrieval module based on the powerful multimodal large language model. Moreover, to assess the generalization of current multimodal sarcasm detection models, we also propose a new OOD test set, RedEval. Experimental results demonstrate that our method is effective and outperforms previous baselines by a large margin, achieving the SOTA performance on both in-domain MMSD2.0 and out-of-domain RedEval datasets.

Limitations

Our method is constrained by the foundational performance of LLM, the visual encoder, and the adapter themselves. Additionally, for the pure LLM-based methods, the quality of the image captions used as visual information input also limits the final performance of the model.

Ethics Statement

We affirm that our work here does not exacerbate the biases already inherent in the large language models. The dataset we crawled is sourced from the official public interfaces of Reddit, which met the requirements. The dataset is only for academic research. The quality of our dataset was confirmed by graduate students at Chinese universities who were paid properly.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) No. U23B2052, Program for Youth Innovative Research Team of BUPT No. 2023QNDT02.

References


Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,


A Instruction Template for LLMs

- **Pure LLM:** “Please select the sarcasm label of ‘<sample text>’ from {0,1}.”

- **LLM with Image Captions:** “Please select the sarcasm label of ‘<sample text ### sample image caption>’ from {0,1}.”

- **LLM with Demonstrations:** “Here is a demonstration: ‘<demonstration text ### demonstration image caption>’, label: ‘<demonstration label>’. Based on the above demonstration, please select the sarcasm label of ‘<sample text ### sample image caption>’ from {0,1}.”