Decoding Sarcasm in Images and Text

Abdullah Mazhar, Aditya Arya,Priyash Shah,
Sanskar Ranjan, Sanya Madan
Indraprastha Institute of Information Technology, Delhi

2nd October 2024

Abstract

Decoding sarcasm in multimodal content, including **images** and **text**, is a challenging task requiring advanced reasoning and integration of diverse modalities. This paper presents a comprehensive approach leveraging **state-of-the-art** language and vision models for sarcasm explanation. Our work combines **common sense reasoning** generated by models like **LLaVA** with textual and visual features processed through fine-tuned models such as **BART**, **GEMMA**, and **Phi 3 Mini**. By enhancing the **MORE dataset** with common sense reasoning, we provide enriched contextual information that significantly improves the quality of sarcasm explanations. Our results demonstrate the effectiveness of **few-shot prompting** and **fine-tuning techniques** across evaluation metrics like **BLEU**, **ROUGE**, and **METEOR**, setting new benchmarks in multimodal sarcasm detection and explanation. This study highlights the critical role of integrating **multimodal data** and **task-specific strategies** to enhance reasoning and explanation generation, paving the way for broader applications in sarcasm understanding.

1 Introduction

Decoding sarcasm from multimodal content, such as images and text, has seen significant advancements. Our work builds upon these developments, leveraging state-of-the-art models and techniques for sarcasm explanation generation. Initially, we compared **Llama 3.2** with **Llama 3.1** using both zero-shot and few-shot methods, alongside existing cross-modal attention-based architectures.

To enhance common-sense reasoning from images, we employed **LLaVA** and **Llama 3.2**. The reasoning outputs were combined with text captions and further processed by **Llama 3.1** and other language models for generating final sarcasm explanations. Our experiments consistently demonstrated improvements in few-shot settings across standard evaluation metrics such as BLEU, ROUGE, and METEOR.

In addition, we fine-tuned multiple models, including **BERT**, **Gemma**, **Llama 3.1**, **Zephyr-SFT**, and **Phi-3 mini**, for multimodal sarcasm detection and explanation. Among these **Phi-3 mini** achieved the best performance, setting new benchmarks on our enhanced **MORE** dataset. This dataset was augmented with common-sense reasoning generated by **LLaVA**, further enriching the contextual information available for sarcasm explanation.

Our work highlights the effectiveness of integrating fine-tuned language models and advanced multimodal architectures to achieve superior sarcasm reasoning and explanation. Moving forward, we aim to refine these models further and explore broader applications of multimodal sarcasm detection.

1.1 Motivation

Sarcasm detection and reasoning remain challenging tasks in natural language understanding due to their reliance on contextual nuances, hidden meanings, and subjective interpretation. While

current models have shown progress in detecting sarcasm, they often fail to provide effective explanations that capture the intricate interplay between textual, visual, and contextual cues.

The absence of robust sarcasm explanations limits the applicability of these models in real-world scenarios, such as improving user interactions on social media or enhancing conversational agents. Incorporating common sense reasoning, derived from visual elements and their corresponding captions, can bridge this gap by providing a more grounded understanding of sarcasm. This integration not only aids in better sarcasm detection but also enriches explanation generation, making models more interpretable and reliable.

Our work is motivated by the need to enhance sarcasm explanation through a multimodal approach, combining advances in common sense reasoning, caption analysis, and fine-tuned language models. By addressing these challenges, we aim to contribute to the development of models capable of reasoning about sarcasm with greater depth and accuracy.

1.2 Dataset Augmentation

We continue to utilize the MORE dataset presented in [1] and are enhancing it with additional multimodal examples from social media, focusing on image-text pairs with sarcasm. We are adding a new column called common sense reasoning (from LLaVA) to the MORE dataset to make a new extended More dataset.

1.3 Novelty

Our work introduces a novel approach to sarcasm explanation by leveraging common sense reasoning (CSR) to reduce reliance on visual inputs while maintaining state-of-the-art performance. Unlike previous models that depend heavily on multimodal data, including images, our framework achieves superior sarcasm explanation using only textual inputs and CSR. This significantly reduces computational overhead and broadens the applicability of sarcasm explanation models to text-only scenarios.

By integrating CSR derived from models like **LLaVA** and combining it with fine-tuned language models such as **Phi-3 mini** and **Zephyr-SFT**, our approach captures nuanced sarcasm reasoning and delivers explanations that align closely with human interpretation. This innovation ensures that the model can operate effectively in settings where visual data is unavailable, setting a new benchmark for multimodal and text-only sarcasm explanation.

Our results demonstrate that this streamlined methodology not only enhances efficiency but also surpasses existing models in metrics like BLEU, ROUGE, and METEOR. This contribution establishes a new paradigm for sarcasm explanation that is both resource-efficient and highly accurate.

1.4 Strategy and Experiments

We have been testing **Llama 3.2** to extract common sense reasoning from multimodal content (images and text) and using **Llama 3.1** and **Quen2.5** to generate sarcasm explanations. First, **Llama 3.2** generates common sense reasoning from the sarcastic image. Then, we combine the image description with the text caption and feed them into a text generation model to produce the final sarcasm explanation.

The process involves the following steps:

- Step 1: Use **Llama 3.2** to generate an image description and extract common sense reasoning from the image.
- Step 2: Combine the image description with the text caption.
- Step 3: Feed the combined inputs into a text generation model (**Llama 3.1** or **Quen2.5**) to generate the final sarcasm explanation.

Images with Common Sense Reasoning using LLaVA

Image 1

Common Sense Reasoning: The meme shows a white car parked in a handicapped space, creating frustration for individuals needing the space.....

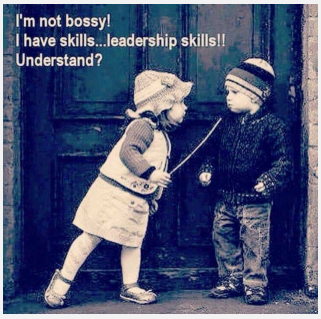

Image 2

Common Sense Reasoning: Two children are seen playing, one asserting leadership skills humorously by saying, "I'm not bossy."....

Above are two of the common sense reasoning examples generated from LLaVA model.

1.5 Reproducibility:

The code for this project is available on GitHub. For details on implementation, datasets, and to reproduce our results, visit our repository: `https://github.com/plon-Susk7/Decoding-Sarcasm`

2 Related Work

Sarcasm explanation in multimodal contexts has gained significant attention in recent years, with numerous studies exploring the integration of textual, visual, and audio cues for a deeper understanding of sarcastic expressions. Below, we discuss key contributions in this domain.

Desai et al. (2022) Desai et al. introduced the MuSE task (Multimodal Sarcasm Explanation), aimed at explaining sarcasm in social media posts that combine text and images. They proposed the ExMore model, a transformer-based encoder-decoder architecture leveraging cross-modal attention to capture incongruities between textual and visual elements, which are central to sarcasm detection.

Hasan et al. (2023) Hasan et al. focused on sarcasm explanation in multimodal, multi-party dialogues by proposing the Multimodal Aware Fusion (MAF) framework. This framework integrates textual, audio, and visual cues into a BART-based architecture to detect and explain sarcasm. They employed the WITS dataset, a rich corpus of multimodal dialogues with manually annotated sarcastic utterances, enabling deeper contextual analysis.

TEAM (2023) The TEAM framework introduced by TEAM (2023) addressed the challenge of generating detailed sarcasm explanations in multimodal posts by combining textual and visual data with external knowledge. By incorporating multi-source semantic graphs and leveraging external knowledge bases such as ConceptNet, the framework enhanced sarcasm reasoning and explanation generation.

EDGE (2024) The EDGE model proposed by EDGE (2024) advanced sarcasm explanation in dialogues by integrating sentiment analysis alongside textual, audio, and visual cues. Using sentiment-enhanced context encoding, the model improved sarcasm understanding by capturing emotional tones from multiple modalities. This sentiment-aware approach resulted in a more nuanced explanation of sarcastic utterances.

These works collectively demonstrate the potential of multimodal approaches, integrating diverse modalities such as text, image, audio, and external knowledge, to improve sarcasm detection and explanation. Our work builds on these insights, exploring additional dimensions for enhanced sarcasm reasoning.

3 Methodologies

We employed a range of vision-based and text-based models to decode sarcasm in multimodal inputs. Below are the detailed methodologies for our approach:

3.1 Common Sense Reasoning

We are using vision transformer models to extract information that can be used to understand the situation shown in the image. This information termed as **Common Sense Reasoning** extracted from the image contains the information about cause-effect of the situation shown in the sarcastic image and figurative understanding consisting of different types of deeper meaning shown in the picture. Below is the promp that was used to extract the common sense reasoning from the images.

Prompt for Extracting Common Sense Reasoning

Instruction: Analyze the following sarcastic meme image to extract common sense reasoning. These relationships should capture the following elements:
1. Cause-Effect: Identify concrete causes or results of the situation depicted in the meme.
2. Figurative Understanding: Capture underlying metaphors, analogies, or symbolic meanings that convey the meme's deeper message, including any ironic or humorous undertones.

3.2 Cross-attention Fine-tuning Architecture :

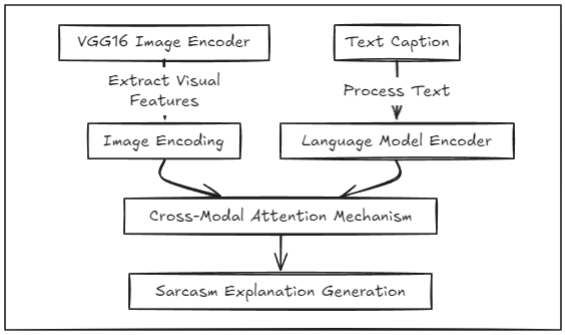

Figure 1: Cross-Attention Architecture

3.2.1 BART, GEMMA, and PHI 3.0 Mini Instruct: cross attention finetuning

For BART, GEMMA, and PHI 3.0 Mini Instruct, we implemented a multimodal architecture based on Desai et al. (2021) [**?**]. This approach integrates visual and textual information processing:

- **Image Encoding:** Utilizes the VGG16 convolutional neural network, pre-trained on ImageNet, to extract high-level visual features from the input images.
- **Text Processing:** Employs the respective language model (BART, GEMMA, or PHI 3.0) to process the textual component of the input.
- **Cross-Modal Attention:** Implements a cross-attention mechanism to fuse the visual and textual features, allowing the model to attend to relevant image regions while generating text.
- **Sarcasm Explanation Generation:** The fused multimodal representation is then used to generate a coherent explanation of the sarcasm present in the input.

3.2.2 Llama 3.1 and Quen2.5: One shot and few shot with CSR from LLaVA

For Llama 3.1 and Quen2.5, we used a two-step methodology to generate sarcasm explanations based on image descriptions:

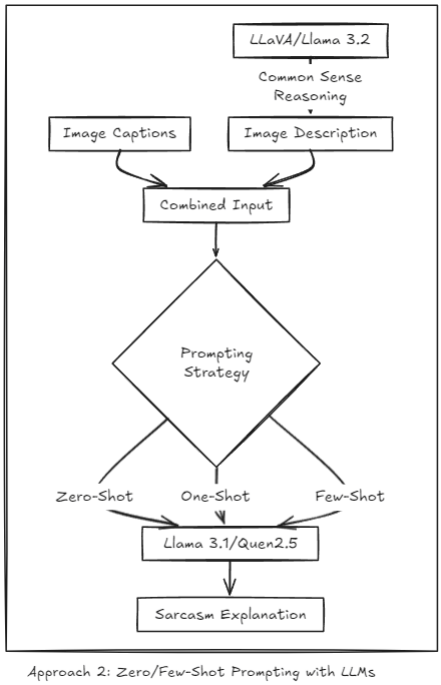

Figure 2: Prompting Mechanism Architecture

1. **Common sense reasoning generation:** The multimodal Llama 3.2 model is first employed to generate detailed descriptions of the input images.

2. **Sarcasm Explanation:** These image descriptions are then combined with the original text input and fed into Llama 3.1 or Quen2.5. Using carefully crafted zero-shot and few-shot prompting methods, Llama 3.1 and Quen2.5 generate explanations of the sarcasm present in the multimodal input.

3.3 Fine-tuning with Common Sense Reasoning from LLaVA

To enhance sarcasm explanation, we fine-tuned BART and GEMMA models on a text-to-text generation task, using common sense reasoning from LLaVA. The process involved:

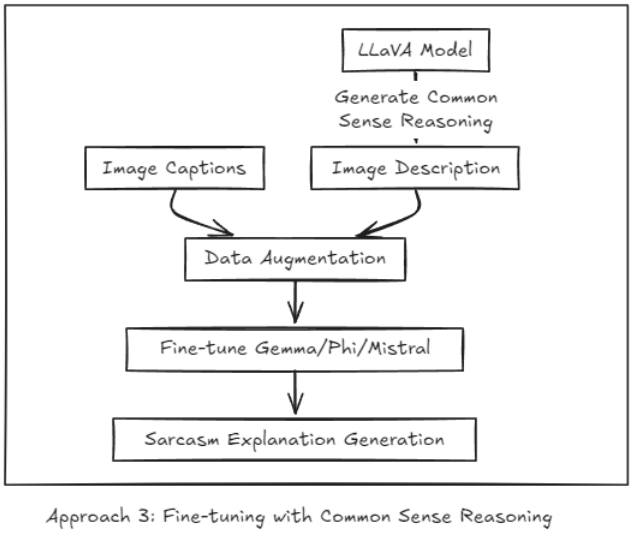

Figure 3: Fine-Tuning Architecture

- **Data Augmentation:** Enhanced the MORE dataset with LLaVA-generated common sense reasoning combined with image captions as input to generate sarcasm explanations.
- **Text Generation Fine-tuning:** BART and GEMMA were fine-tuned to generate sarcasm explanations by learning from both captions and LLaVA's reasoning.

This approach improved sarcasm explanation quality by leveraging enriched context through LLaVA's common sense reasoning in text format.

4 Results and Discussion

In this section, we present the quantitative results of the models we experimented with. Both zero-shot and few-shot scenarios were evaluated for sarcasm explanation generation, using various metrics such as BLEU, ROUGE, and METEOR.

The comparison across models highlights that the few-shot settings consistently perform better than zero-shot across all metrics. Notably, **Llama 3.2 for common sense reasoning** in combination with **Llama 3.1** and **Quen2.5** as language models outperformed in several key areas, including BLEU-3, BLEU-4, and ROUGE-L scores. This shows the effectiveness of combining vision models with language models for sarcasm understanding.

4.1 Quantitative Results

4.1.1 Cross-attention Finetuning

The following table summarizes the performance metrics for baseline models (BART, GEMMA, PHI 3.0 Mini) that were used to generate sarcasm explanations:

Table 1: Performance Metrics for Sarcasm Explanation Models (Baseline Models)

Metrics	BART	GEMMA	PHI 3.0 Mini
ROUGE-1 F1	0.27	0.20	0.19
ROUGE-2 F1	0.12	0.06	0.05
ROUGE-L F1	0.25	0.16	0.16
METEOR	0.27	0.36	0.33
BLEU Score	0.05	0.35	0.36
BERTScore F1	0.78	0.06	0.10

4.1.2 Zero-shot and few shot prompting

The following table summarizes the results using **Llama 3.2 for common sense reasoning** and **Llama 3.1** and **Quen2.5** as language models, in both zero-shot and few-shot settings:

Table 2: Performance of Llama 3.2 (Common Sense Reasoning) with Llama 3.1 and Quen2.5 in Zero-Shot and Few-Shot Settings

Metrics	Llama 3.1 Zero-Shot	Llama 3.1 Few-Shot	Quen2.5 Zero-Shot	Quen2.5 Few-Shot
BLEU-1	0.10	0.12	0.09	**0.12**
BLEU-2	0.03	0.05	0.18	**0.20**
BLEU-3	0.01	0.02	0.28	**0.30**
BLEU-4	0.01	0.01	0.36	**0.38**
ROUGE-1 F1	0.22	**0.24**	0.19	0.23
ROUGE-2 F1	0.06	**0.08**	0.05	0.06
ROUGE-L F1	0.17	**0.20**	0.15	0.18
Simplified METEOR	0.37	**0.42**	0.35	0.39

4.1.3 Fine-tuning with Common Sense Reasoning

We experimented with fine-tuning different LLMs using the common sense reasoning from the image augmented into the MORE dataset to get sarcasm explanation.

Table 3: Performance of Models on Sarcasm Explanation Evaluation with Common Sense Reasoning

Metric	Llama 3.1 (8B)	phi 3 mini (4k)	Mistral-Nemo	Zephyr-SFT	Llama 3.2 Turbo	Llava
ROUGE-1	0.35	**0.49**	0.37	**0.49**	0.48	0.03
ROUGE-2	0.21	**0.31**	0.24	**0.31**	**0.31**	0.00
ROUGE-L	0.30	**0.46**	0.35	0.45	0.45	0.03
BERT Precision	0.87	**0.92**	0.87	0.89	0.91	0.52
BERT Recall	**0.91**	**0.91**	0.90	**0.91**	**0.91**	0.56
BERT F1	0.88	**0.91**	0.88	0.90	**0.91**	0.54
METEOR	0.45	0.47	0.48	**0.51**	0.49	0.05
BLEU	0.13	**0.27**	0.16	0.22	0.24	0.00

5 Result Analysis

The evaluation highlights the strengths of fine-tuned models and few-shot prompting methods for sarcasm explanation. Among the baseline models, BART performed well in ROUGE metrics but lagged in fluency and relevance compared to GEMMA, which excelled in BLEU scores. PHI 3.0 Mini showed balanced but slightly weaker performance overall.

Few-shot prompting outperformed zero-shot prompting, with Quen2.5 in the few-shot setting achieving the highest BLEU-4 score (0.38), showcasing its ability to generate high-quality explanations.

Fine-tuning with common sense reasoning proved highly effective, with Phi 3 Mini and Zephyr-SFT achieving top ROUGE and METEOR scores. Larger models like Llama 3.2 Turbo performed consistently well, while Llava struggled, highlighting its limitations for this task.

In summary, fine-tuning multimodal models with common sense reasoning outperformed both baseline models and prompting-based approaches, emphasizing the importance of leveraging task-specific strategies for nuanced sarcasm explanation.

5.1 Evaluation Metrics Used

To comprehensively assess the performance of each model, we utilized a diverse set of evaluation metrics:

- **ROUGE Scores (1, 2, and L):** Measuring the overlap of n-grams between generated explanations and human references.
- **METEOR:** Evaluating the semantic similarity between generated and reference explanations.
- **BLEU Score:** Assessing the quality of generated explanations based on n-gram precision.
- **BERTScore:** Computing the semantic similarity using contextual embeddings.

6 Conclusion

This project presents an extensive exploration of multimodal sarcasm explanation using advanced language and vision models. By leveraging common sense reasoning extracted through LLaVA and integrating it with fine-tuned models like BART, GEMMA, and Phi 3 Mini, we significantly improved the quality of sarcasm explanations. The results demonstrate the superiority of few-shot prompting and fine-tuning approaches, especially when enriched with common sense reasoning, over traditional baseline models. Our work also highlights the critical role of multimodal data and task-specific strategies in achieving robust sarcasm detection and explanation. Moving forward, we aim to refine our methodologies further, enhance the dataset, and explore broader applications of multimodal reasoning frameworks.

7 Video Demonstration

We have a made a web application and a demo of it is available on this link.

Access the document here

The code for this project is available on GitHub. For details on implementation, datasets, and to reproduce our results, visit our repository: `https://github.com/plon-Susk7/Decoding-Sarcasm`

References

[1] Desai, P., Chakraborty, T., & Akhtar, M. S. (2022). Nice perfume. How long did you marinate in it? Multimodal Sarcasm Explanation. In Proceedings of the AAAI Conference on Artificial Intelligence, 36, 10563–10571.

[2] Babanejad, N., Davoudi, H., An, A., & Papagelis, M. (2020). Affective and contextual embedding for sarcasm detection. In Proceedings of the 28th International Conference on Computational Linguistics, 225–243.

[3] Bedi, M., Kumar, S., Akhtar, M. S., & Chakraborty, T. (2021). Multi-modal sarcasm detection and humor classification in code-mixed conversations. IEEE Transactions on Affective Computing, 14(2), 1363–1375.

[4] Bouazizi, M., & Ohtsuki, T. O. (2016). A pattern-based approach for sarcasm detection on Twitter. IEEE Access, 4, 5477–5488.

[5] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524.

[6] Kumar, S., Kulkarni, A., Akhtar, M. S., & Chakraborty, T. (2022). When did you become so smart, oh wise one?! Sarcasm explanation in multi-modal multi-party dialogues. arXiv preprint arXiv:2203.06419.

[7] Lewis, M. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[8] Mishra, A., Tater, T., & Sankaranarayanan, K. (2019). A modular architecture for unsupervised sarcasm generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 6144–6154.

[9] Peled, L., & Reichart, R. (2017). Sarcasm sign: Interpreting sarcasm with sentiment-based monolingual machine translation. arXiv preprint arXiv:1704.06836.

[10] Qiao, Y., Jing, L., Song, X., Chen, X., Zhu, L., & Nie, L. (2023). Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 37, 9507–9515.

[11] Schifanella, R., De Juan, P., Tetreault, J., & Cao, L. (2016). Detecting sarcasm in multimodal social platforms. In Proceedings of the 24th ACM International Conference on Multimedia, 1136–1145.

[12] Tay, Y., Tuan, L. A., Hui, S. C., & Su, J. (2018). Reasoning with sarcasm by reading in-between. arXiv preprint arXiv:1805.02856.