DA421M Course Project

Sahil Danayak[1] and Himangshu Deka[2]

[1] Roll No. 210101092,Dept. Of Computer Science and Engineering, IIT Guwahat
[2] Roll No. 210101050, Dept. Of Computer Science and Engineering, IIT Guwahati

Abstract. Sarcasm is often challenging to interpret due to the discrepancy between literal and intended meanings, where understanding the real intent requires contextual knowledge. This is especially critical in today's social media landscape, where sarcasm detection has become essential in mitigating the potential negative impacts, such as cyberbullying. Traditional approaches have focused on textual analysis, yet the rise of multimedia communication shows that effective sarcasm detection needs a multi-modal approach. This review delves into existing research on sarcasm detection utilizing multi-modal data and presents an evaluation of these models based on accuracy and F1-score.
To address the limitations of current methods, we introduce a novel approach, an innovative framework for multi-modal sarcasm detection. Our model integrates our improved CLIP module to capture nuanced text-image interactions by embedding cross-modal context directly within each encoder. Additionally, we propose a mechanism that leverages a dynamic memory channel to retain crucial samples during inference, functioning as a non-parametric classifier. Our proposal achieves new benchmarks on MMSD2.0, with a 1.08% boost in accuracy and a 1.51% improvement in F1-score and a generic trend of improvement in Recall and Accuracy.

Keywords: CLIP · MMSD2.0

1 Introduction

Detecting sarcasm on social media is essential for applications like sentiment analysis, product reviews, and user-generated content moderation, where understanding the intended message is critical to improve user experience and prevent miscommunication. Traditional sarcasm detection models largely rely on identifying linguistic cues in text, but these often fall short, especially on platforms where images and other non-verbal elements contribute significantly to the message's meaning. As multi-modal communication becomes more prevalent, text-only methods risk missing the subtle interplay between verbal and visual sarcasm cues.

Despite progress in multi-modal sarcasm detection, recent studies (Xu, Zeng, and Mao 2020; Pan et al. 2020; Liang et al. 2021, 2022) in this field, Qin et al. (2023) show that popular benchmarks, such as MMSD[Multi Model Sarcasm detection](Cai, Cai, and Wan 2019) , may contain unintended artifacts, leading models to pick up on spurious patterns rather than true sarcasm indicators. To address these issues, MMSD2.0 has been introduced, refining these benchmarks to minimize data biases and revealing that multi-modal sarcasm detection still faces significant challenges.

In this context, we propose to enhance sarcasm detection by integrating an Interactive Contrastive-Language Image Pre-Training(CLIP) with a mechanism that leverages a dynamic memory channel to retain crucial samples during inference, functioning as a non-parametric classifier . This approach improves text-image representation by embedding cross-modal context directly into each encoder, allowing the model to capture complex interactions between text and visual data while simultaneously ensuring that it retains dynamic, valuable test sample representations, using this memory as a non-parametric classifier for more robust sarcasm recognition.

- We introduce a framework for dependable multi-modal sarcasm detection. This framework utilizes an improvised CLIP module to enhance text-image representation by embedding cross-modal information within each encoder, combined with a dual channel memory module for improved sarcasm detection accuracy and resilience.
- We design an optimized training approach that maximizes computational efficiency and model performance.
- Comprehensive testing on the refined MMSD2.0 benchmark confirms that our proposal in some cases surpasses previous methods in F1 scores and other metrics.

2 Related Works

In recent years, the task of multi-modal sarcasm detection has gained traction as researchers have explored the interplay of textual and visual cues. Early work on sarcasm detection was rooted in textual analysis alone (Bouazizi and Ohtsuki, 2015; Amir et al., 2016; Baziotis et al., 2018), leveraging linguistic features and sentiment analysis to identify sarcasm. However, with the rise of social media platforms that combine text and image in a single post, purely text-based approaches began facing limitations, as visual context often plays a critical role in conveying sarcastic tones.

SchiFanella et al. (2016) were among the first to integrate text and images, using multi-modal social media posts to capture sarcasm cues across both text and visual elements. This pioneering work marked a significant step forward in understanding how humor and sarcasm are conveyed across modalities, prompting subsequent research on multi-modal fusion methods. In 2019, Cai, Cai, and Wan advanced the field by introducing the MMSD benchmark, which established a hierarchical fusion model to better integrate text with both images and image attributes. Using a pre-trained ResNet model for regional image vectors and GloVe embeddings for high-level image features, they structured their model with two layers of fusion: an early fusion that initialized text representations with image attribute vectors, and a representation fusion that integrated raw and guided vectors from each modality. This model effectively leveraged cross-modal interactions to achieve improved performance in sarcasm classification, and MMSD soon became a foundational benchmark, catalyzing additional research (Xu, Zeng, and Mao, 2020; Pan et al., 2020; Liang et al., 2021, 2022; Liu, Wang, and Li, 2022; Qin et al., 2023; Wen, Jia, and Yang, 2023; Tian et al., 2023; Wei et al., 2024).

While MMSD provided a strong baseline, its reliance on certain spurious cues posed risks of model bias. This limitation was addressed in Qin et al. (2023), who introduced MMSD2.0, a refined version that removed these spurious cues and corrected mislabeled samples. Their analysis revealed a considerable performance drop when re-evaluating state-of-the-art methods on MMSD2.0, indicating that many models were inadvertently relying on misleading information. This insight led Qin et al. to call for more stable approaches to multi-modal sarcasm detection. Our work builds upon this with the Improvised CLIP module enhanced with a dual channel memory framework, aiming to mitigate these biases while advancing reliability in sarcasm detection.

Further enhancing multi-modal approaches, the Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021) has demonstrated strong potential across vision-language tasks. CLIP's ability to align visual and textual information in a shared embedding space has made it a popular foundation for domain-specific adaptations, leading to substantial improvements in various applications. Li, Shakhnarovich, and Yeh (2022) adapted CLIP for phrase localization, enabling it to recognize contextually appropriate regions in images. Building on CLIP's architecture, Zhou et al. (2022) introduced Context Optimization (CoOp) to enhance CLIP's adaptability for downstream tasks by automating prompt engineering. Instead of relying on manual prompts, CoOp learns context vectors that act as optimized prompts for specific tasks. It offers two styles: unified context, which uses the same context vectors across all classes, and class-specific context, which tailors vectors for each class

Liang et al. (2023) extended CLIP's capabilities to open-vocabulary semantic segmentation, broadening its interpretive scope to include unseen categories. Similarly, Wang et al. (2023) applied CLIP for action recognition, making it possible to predict actions by linking textual and visual cues. Building on these advances, Ganz et al. (2024) enhanced CLIP by embedding text into the visual encoder, further improving cross-modal alignment. Inspired by Ganz et al., our work advances this adaptation by conditionally embedding both text and images into their respective encoders, thereby capturing deeper semantic relationships. Unlike Ganz et al., who focused exclusively on embedding text into the visual encoder, our approach also embeds images into the text encoder, facilitating a bi-directional understanding necessary for capturing nuanced sarcasm across modalities. Additionally, our work diverges from general classification, addressing the unique complexities of multi-modal sarcasm detection.

Finally, memory-enhanced prediction models have gained attention as a promising approach for improving model performance over time by leveraging cognitive-inspired architectures. Drawing on the foundations of cognitive science (Stokes, 2015; Baddeley, 2000), memory mechanisms were first introduced into neural networks by Weston, Chopra, and Bordes (2014) and Sukhbaatar et al. (2015), enabling networks to retain critical information over multiple processing steps. This concept was subsequently applied in several studies (Wu et al., 2018; Wen, Jia, and Yang, 2023), where memory modules helped improve model training through the retention of relevant past data. Recently, Zhang et al. (2024) and Wei et al. (2024) utilized memory-based approaches to store historical information, thereby enhancing predictive accuracy by referencing prior knowledge. Our approach integrates a dynamic memory-enhanced predictor within the Improvised CLIP module enhanced with a dual channel memory framework to address the need for sarcasm detection models that can adapt to changing contexts. Unlike previous methods, our memory component is designed to update during testing, allowing the model to adjust to new instances and employ relevant historical information dynamically. This approach promotes improved accuracy and robustness by preserving critical multi-modal patterns across various scenarios.

Each of these studies represents a unique stride toward the ultimate goal of precise sarcasm detection in multi-modal contexts, highlighting the complexity of capturing semantic contradictions and aligning meaning across text and images. Our proposed framework synthesizes insights from hierarchical fusion, CLIP adaptation, and memory-enhanced prediction, contributing a comprehensive model to the growing field of multi-modal sarcasm detection.

3 Methods

3.1 Improvised CLIP Module

Our Improvised CLIP Module processes paired inputs of text and images, aiming to create an interactive, multi-modal representation that captures intricate relationships between visual and textual cues. Here's how it works in detail:

Encoding Text and Image Inputs The text encoder processes the input text, breaking it down into tokens and encoding each one into a high-dimensional vector. Special tokens, such as beginning-of-sequence (BOS) and end-of-sequence (EOS) markers, are also added to indicate the start and end of the input text. The vision encoder, meanwhile, processes the image by dividing it into patches, each of which is encoded into a vector. A special classification (CLS) token is added to represent the entire image. Reducing Complexity with Self-Attention Layers: To manage computational complexity, the Improvised CLIP Module only uses the top-n self-attention layers from each encoder, a strategy designed to balance efficiency and performance. This approach retains the critical information required for sarcasm detection without overwhelming computational resources.

Interactive Representation After obtaining the encoded features from both text and image, the Improvised CLIP Module combines them using concatenation. This step generates an interactive representation that captures relationships between the image and text, potentially highlighting cues that could indicate sarcasm. This interactive representation allows the model to interpret text and image jointly, rather than in isolation, which is particularly useful for detecting sarcasm where visual or textual context alone may not be sufficient.

3.2 Memory-enhanced prediction:Dual Channel Memory Module

Inspired by cognitive science (Stokes 2015; Baddeley 2000), memory has been introduced to enhance neural networks (Weston, Chopra, and Bordes 2014; Sukhbaatar et al. 2015). Several studies (Wu et al. 2018; Wen, Jia, and Yang 2023) have used memory mechanisms to improve model training, and some (Zhang et al. 2024; Wei et al. 2024) leverage memory to store historical knowledge, enhancing prediction accuracy. In this work, we introduce a memory-enhanced predictor for multi-modal sarcasm detection. In contrast to other methods, our memory dynamically updates during testing, utilizing relevant historical information for improved accuracy and robustness.

The Memory-Enhanced Predictor builds on the Improvised CLIP Module's representations, adding a mechanism to incorporate relevant past interactions. The dual channel memory module provides the model with "memory," enabling it to reference historical data when predicting sarcasm. This memory component is key to improving prediction accuracy in sarcasm detection, where context over time can greatly influence interpretation.

Memory Module The memory module has a memory bank that stores feature representations from previously seen examples. This memory serves as a repository of historical information, which the model can access during prediction.

Dynamic Updates During the testing phase, memory module's memory updates dynamically with each new example, enabling the model to continuously refine its historical knowledge as it encounters more data. This dynamic updating allows the model to adapt and incorporate recent interactions into its predictions. Leveraging Historical Information: By accessing the stored historical features, the model can compare a current input against similar past instances, aiding in the detection of sarcastic tone that may rely on contextual patterns established in prior examples. For instance, sarcasm is often subtle and may depend on recurring themes or previously encountered sentiments, which memory module's memory can help capture.

3.3 Training Strategy

To train the Improvised CLIP Module effectively, the model uses a combined loss function that balances classification accuracy and feature alignment:

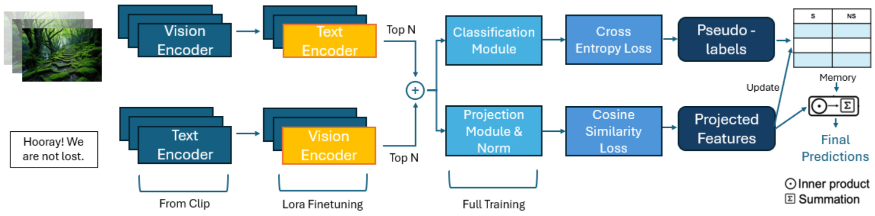

Fig. 1. Architecture Overview: Training Improvised CLIP Module : Vision and text representations are obtained through distinct encoders, which are then integrated into the top-n layers of the counterpart modality's encoder for enhanced interaction. The top-n layers undergo fine-tuning using LoRA, while the remainder of the encoder remains fixed. The resulting vision and text representations are concatenated and utilized to train a classification module aimed at detecting sarcasm. Additionally, a projection module is trained to map these representations into a latent space. Dual Channel Memory : During the inference phase, Improvised CLIP Module creates interactive representations. The classification module assigns pseudo-labels, while the projection module generates projection features. This part of the architecture enriches its dynamic memory with these features and pseudo-labels. The final prediction is achieved by comparing the projected features of the current sample with those stored in memory.

Cross-Entropy Loss This loss function is used for sarcasm classification, guiding the model to correctly categorize each instance as either sarcastic or non-sarcastic. Cross-Entropy Loss penalizes misclassifications, helping the model to distinguish sarcastic instances from non-sarcastic ones.

Cosine Similarity Loss Cosine Similarity Loss is applied to ensure that similar samples have closely aligned representations in the feature space. This loss function encourages samples with similar sarcasm cues to be grouped together, while dissimilar samples are positioned further apart. By optimizing for Cosine Similarity Loss, the model is better able to capture nuanced relationships and patterns in the data, which aids in sarcasm detection where such similarities can be subtle. The combination of these two losses helps the Improvised CLIP Module-dual channel memory module learn both to classify sarcasm accurately and to organize features in a way that reflects the underlying relationships between different samples.

3.4 Summary

The Improvised CLIP Module-Dual Channel framework leverages both interactive multi-modal representation and memory-based context-awareness to tackle the complexities of sarcasm detection. Our Improvised CLIP Module provides joint representations of text and image, enabling the model to capture sarcasm-related cues that span both modalities. The Dual Channel Memory Module, on the other hand, enriches this approach by dynamically incorporating historical information through its memory module. This allows the model not only to consider the immediate content of each input but also to reference past instances, which can provide crucial context for correctly identifying sarcasm. Together, Improvised CLIP Module enforce with a Dual Channel Memory module form a robust framework for multi-modal sarcasm detection, benefiting from both present and past context.

4 Experiments

4.1 Datasets

We use the MMSD dataset developed by Cai and Wan (2019) along with its enhanced version, MMSD2.0 (Qin et al., 2023), both of which serve as important benchmarks in multimodal sarcasm detection.

The datasets are accessible at:
https://github.com/JoeYing1019/MMSD2.0/tree/main/data

4.2 Metrics

We use accuracy (Acc.), precision (P), recall (R), and F1-score (F1) as metrics to evaluate the performance of our model.

4.3 Baselines

We compare the effectiveness of the Improvised CLIP Module with Dual Channel Memory module framework against several unimodal and multi-modal methods. For text modality methods, we compare with TextCNN (Kim 2014), Bi-LSTM (Graves and Schmidhuber 2005), SMSD (Xiong et al. 2019), and RoBERTa (Liu et al. 2019). For image modality methods, we compare with ResNet (He et al. 2016) and ViT (Dosovitskiy et al. 2021). For state-of-the-art multi-modal methods, we compare with the following:

- **HFM** (Cai, Cai, and Wan 2019): A hierarchical fusion model integrating text and image features.
- **Att-BERT** (Pan et al. 2020): A model using an attention mechanism to capture sarcasm incongruity.
- **CMGCN** (Liang et al. 2022): A method that constructs a cross-modal graph and employs a graph convolutional network.
- **HKE** (Liu, Wang, and Li 2022): A hierarchical framework utilizing multi-head cross-attention and graph neural networks.
- **DIP** (Wen, Jia, and Yang 2023): A dual incongruity perceiving network that extracts sarcastic information from factual and affective levels.
- **DynRT** (Tian et al. 2023): A dynamic routing transformer network for multi-modal sarcasm detection.

- **Multi-view CLIP** (Qin et al. 2023): A method adapted to CLIP for capturing multi-view sarcasm cues.
- **G2SAM** (Wei et al. 2024): A method using global graph-based semantic awareness and leveraging historical knowledge from training samples for multi-modal sarcasm detection.

4.4 Results

To assess the effectiveness of our Improvised CLIP Module-with Dual Channel Memory module framework, we conducted experiments using the original CLIP model as the backbone, referred to as Improvised CLIP Module-with Dual Channel Memory module . In this experiment, we conditioned only the top four layers of the self-attention modules and set the projection dimension d_f to 1024. The LoRA rank r was set to 8, allowing us to fine-tune the self-attention module weight matrices W (specifically W_k, W_v, and W_o). For the with Dual Channel Memory module memory size L, we selected the optimal size from the range $L = \{128, 256, 384, 512, 640, 768, 896, 1024, 1152, 1280\}$.

The main results, displayed in Table 1, show that our framework consistently matches and sometimes betters existing methods, regardless of whether our original CLIP is used as the backbone. This highlights the effectiveness of our training strategy and of our proposal with the Dual Channel Memory module.

Modality	**Method**	**Acc. (%)**	**F1 (%)**	**P (%)**	**R (%)**
Text	TextCNN (Kim 2014)	71.61	69.52	64.62	75.22
Text	Bi-LSTM (Graves and Schmidhuber 2005)	72.48	68.05	68.02	68.08
Text	SMSD (Xiong et al. 2019)	73.56	69.97	68.45	71.55
Text	RoBERTa (Liu et al. 2019)	79.66	76.20	76.74	75.70
Image	ResNet (He et al. 2016)	65.50	57.58	61.17	54.39
Image	ViT (Dosovitskiy et al. 2021)	72.02	69.76	65.26	74.83
Text-Image	HFM (Cai, Cai, and Wan 2019)	70.57	66.87	64.84	69.05
Text-Image	Att-BERT (Pan et al. 2020)	80.03	77.04	76.28	77.82
Text-Image	CMGCN (Liang et al. 2022)	79.83	76.99	75.82	78.01
Text-Image	HKE (Liu, Wang, and Li 2022)	76.50	72.25	73.48	71.07
Text-Image	DIP (Wen, Jia, and Yang 2023)	80.59	78.23	75.52	81.14
Text-Image	DynRT (Tian et al. 2023)	70.37	68.55	63.02	75.15
Text-Image	Multi-view CLIP (Qin et al. 2023)	85.64	84.10	80.33	88.24
Text-Image	G^{2}SAM (Wei et al. 2024)	79.43	78.00	72.04	85.20
Text-Image	Our Proposal ($L = 1024$)	86.10	84.80	84.30	90.45

Table 1. Comparison of Methods by Modality and Performance Metrics v/s our Proposal on MMSD2.0 dataset

Modality	**Method**	**Acc. (%)**	**F1 (%)**	**P (%)**	**R (%)**
Text	TextCNN (Kim 2014)	74.03	72.28	68.30	78.49
Text	Bi-LSTM (Graves and Schmidhuber 2005)	75.04	71.42	69.64	71.33
Text	SMSD (Xiong et al. 2019)	77.42	73.45	72.46	73.76
Text	RoBERTa (Liu et al. 2019)	82.98	79.76	79.86	79.35
Image	ResNet (He et al. 2016)	68.19	59.42	63.89	57.66
Image	ViT (Dosovitskiy et al. 2021)	76.26	73.44	68.95	78.32
Text-Image	HFM (Cai, Cai, and Wan 2019)	73.92	69.94	68.42	72.31
Text-Image	Att-BERT (Pan et al. 2020)	83.04	80.28	80.01	81.62
Text-Image	CMGCN (Liang et al. 2022)	83.43	79.91	79.22	81.47
Text-Image	HKE (Liu, Wang, and Li 2022)	79.13	74.55	77.32	73.53
Text-Image	DIP (Wen, Jia, and Yang 2023)	83.72	80.60	78.96	84.66
Text-Image	DynRT (Tian et al. 2023)	73.64	71.84	67.07	78.59
Text-Image	Multi-view CLIP (Qin et al. 2023)	89.41	88.05	83.16	91.77
Text-Image	G^{2}SAM (Wei et al. 2024)	82.37	81.29	75.89	87.10
Text-Image	Our Proposal ($L = 1024$)	89.95	88.04	86.12	93.24

Table 2. Comparison of Methods by Modality and Performance Metrics v/s our Proposal on MMSD dataset

5 Conclusion

In this study, we presented a novel framework for robust multi-modal sarcasm detection, named Our Proposal. Our approach integrates an Improvised CLIP Module model as the backbone for generating sample representations. Unlike traditional CLIP, our enriches the encoding process by embedding representations from multiple modalities, allowing for deeper capture of complex interactions between text and image. For enhanced inference, we introduced a Memory-Enhanced Predictor that utilizes historical knowledge from previous samples to improve robustness and reliability in sarcasm detection. Our experimental findings confirm that Our Proposal betters upon previous metrics on the MMSD2.0 benchmark.

References

1. Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
2. Xu, N.; Zeng, Z.; and Mao, W. 2020. Reasoning with multi-modal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association. May 2020.
3. Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; and Wang, W. 2020. Modeling Intra and Intermodality Incongruity for Multi-Modal Sarcasm Detection.
4. Liang, et al. 2021. Multi-Modal Sarcasm Detection with Interactive In-Modal and Cross-Modal Graphs.
5. Liang, et al. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network.
6. Liu, H.; Wang, W.; and Li, H. 2022. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement.
7. Qin, L.; Huang, S.; Chen, Q.; Cai, C.; Zhang, Y.; Liang, B.; Che, W.; and Xu, R. 2023. MMSD2.0.
8. Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-Vocabulary Semantic Segmentation With Mask-Adapted CLIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7061–7070.
9. Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In International Conference on Learning Representations.
10. Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, 8748–8763. PMLR.
11. Zhang, Y.; Zhu, W.; Tang, H.; Ma, Z.; Zhou, K.; Zhang, L. 2024. Dual Memory Networks: A Versatile Adaptation Approach for Vision-Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.