

LDGNet: LLMs Debate-Guided Network for Multimodal Sarcasm Detection

Hengyang Zhou^{‡*}
China University of Petroleum
Karamay, China
2021015426@st.cupk.edu.cn

Jinwu Yan[‡]
China University of Petroleum
Karamay, China
2021015387@st.cupk.edu.cn

Yaqing Chen
China University of Petroleum
Karamay, China
2021015435@st.cupk.edu.cn

Rongman Hong
China University of Petroleum
Karamay, China
2021015394@st.cupk.edu.cn

Wenbo Zuo
China University of Petroleum
Karamay, China
2020015542@st.cupk.edu.cn

Keyan Jin^{*}
Macao Polytechnic University
Macao, China
p2317001@mpu.edu.mo

Abstract—Multimodal sarcasm detection aims to uncover the sarcasm emotions expressed through various modalities such as text and image. Previous work has made enlightening exploration in detecting sarcastic sentiments with given domains. However, there remains a gap in utilizing deeper contextual information to capture elusive sarcastic clues, hidden in open-world knowledge such as history, politics, and common sense of life that has not been touched by previous models. To address this gap, a natural idea is to simulate the process of a debate, involving debaters with different viewpoints and judges to collaboratively drive the judgment of emotional expressions. Benefiting from the development of large multimodal language models, and building upon previous advancements, we propose a novel framework called LLMs Debate-Guided Network (LDGNet) for Multimodal Sarcasm Detection. LDGNet effectively leverages large language model debates to uncover subtle emotional information and uses an innovative Judge Network for more reliable and accurate sentiment judgments. Extensive experiments on in-domain and out-of-distribution (OOD) datasets have validated the superiority of our proposed method.

Index Terms—Multimodal sarcasm detection, multimodal debate, LLMs

I. INTRODUCTION

With the advancement of social media technology, an increasing number of users are inclined to express their emotions on social media platforms [1], [2]. The field of Multimodal Sarcasm Detection (MSD) is at the forefront of human sentiment identification, integrating diverse data types such as images and text. Its applications are widespread, notably in healthcare and human-computer interaction [3].

Solving the pain points of MSD tasks mainly stem from two aspects. First, expressed sentiment tendencies from diverse modalities, which increases the complexity of learning. As shown in Fig. 1(a), the text expresses the good weather, and the image expresses the bad rainy weather, which intuitively reflects the conflict between the image and the text.



(a) what a wonderful weather ! Sarcasm



(b) hitting sure has changed in <num> years. Sarcasm

Fig. 1: Examples of sarcasm data from Twitter.

Recent research [4] on multimodal techniques has also referred to the thorny nature of this task. Second, compared to unimodal data, the emotional tendencies expressed in multimodal data are often embedded in a rich trove of open-world knowledge, making them difficult to identify. As Fig. 1(b) shows, both the text and the image convey the meaning of the sports changing, so the conflict between the image and the text cannot be captured. Hindering the layman from judging the user to express sarcastic emotion, it is actually sarcastic emotion in expressing the sport's more complex techniques and devices, losing its original charm and simplicity. Open-world knowledge in areas such as history, politics, and common sense of life is not fully reflected in users' tweets, as it is already possessed by individuals. Unfortunately, previous models have not acquired such open-world knowledge, rendering them unable to learn accurate emotional tendencies from the data, thereby limiting their performance.

Previous MSD works primarily focused on the integration of multi-modal features utilizing diverse approaches. Initial attempts, HFM [2] proposes a multimodal hierarchical framework that fuses text, image, and image attributes. [5], [6] proposes that the pain point of sarcasm detection is the incongruity between modalities. InCrossMGs [7] and CMGCN [8] designs a graph-based approach to capture sarcastic clues. DGP [9] proposed a method to adjust modal weights and bidirectional generation of modal features. HKEmodel [10] proposed atomic-level consistency and composition-level con-

[‡]Both authors contribute equally to this work.

^{*}Corresponding author.

Multimodal Debate Between LLMs

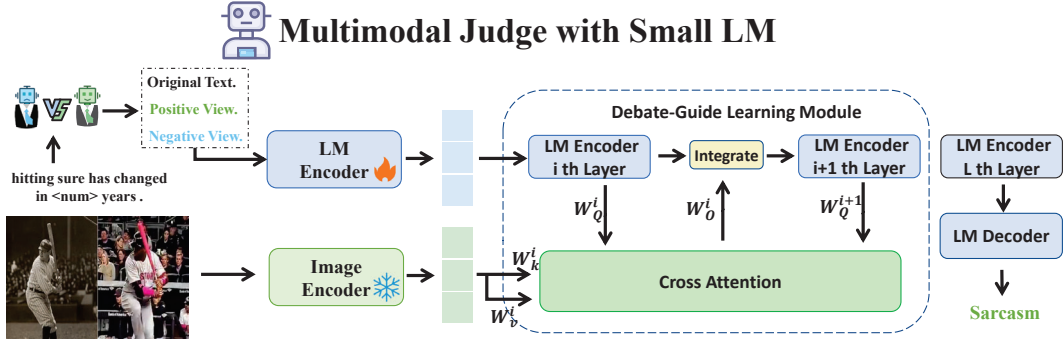
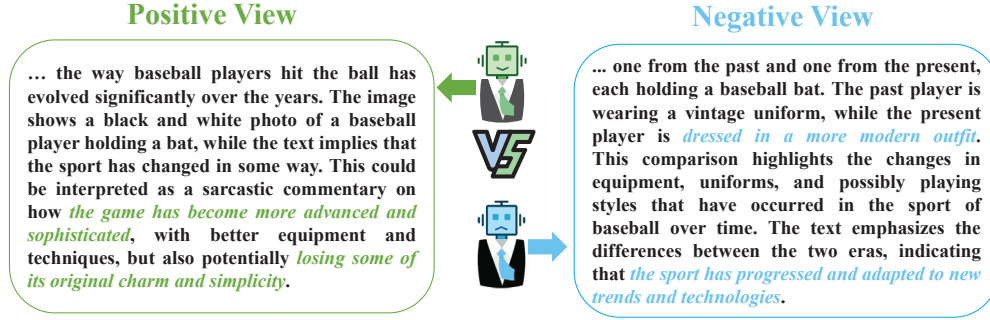


Fig. 2: The overall architecture of our proposed LDGNet.

sistency method modeling methods. MILNet [11] constructs three graphs to learn local and global sarcastic information. Based on the HFM dataset, DMSD-CL [12] proposed an out-of-distribution (OOD) dataset to construct positive samples with different word biases and negative samples with similar word biases. DIP [4] explicitly models emotional incongruity and factual incongruity. G²SAM [13] proposed a fine-grained graph-based model to capture contradictory emotional cues.

Despite advancements in prior research, the improvement of MSD tasks has been hindered by inadequate utilization of open-world knowledge. Due to the existing gap, our motivation is to leverage open-world knowledge in a reasonable manner. Specifically, reasonable open-world knowledge is first generated through debates between large language models (LLMs), so that the engulfed semantic information can be analyzed from different sentiment perspectives. A trustworthy judge network necessitates a comprehensive consideration of pleadings generated by large language models from various sentiment perspectives.

The main contributions can be summarized as follows:

- To our best knowledge, this is the first work that utilizes advanced LLMs to investigate multimodal sarcasm detection from a novel perspective of leveraging open-world knowledge.
- We introduce a novel and insightful approach that leverages multimodal debate among LLMs to acquire open-world knowledge from arguments of diverse emotional expressions, and subsequently determines emotional tendencies through a reliable judge module.

- Extensive experiments conducted on In-Domain and Out-of-Distribution datasets demonstrate that our model outperforms previous state-of-the-art MSD baselines.

II. METHODOLOGY

As illustrated in Figure 2, the proposed LDGNet framework comprises two primary modules: (a) the Large Language Models (LLMs) Debating Module, which prompts different emotional tags on LLMs to generate high-quality open-world knowledge from the original samples through debates; (b) the Multimodal Judge Module, which leverages the open-world knowledge generated by the LLMs Debate Module to make more comprehensive judgments on emotional tendencies. In the following sections, we will elaborate on these two modules.

A. Multimodal Debate Module

Given an image-text pair $\langle I_i, T_i \rangle$, the debate participants are instructed to analyze the pair from the perspective of label set $y_i \in \{\text{positive}, \text{negative}\}$ based on specific tasks using a prescribed prompt template. Each debate participant generates a rationale, and the debate result is expressed as $R_i = \{R_i^{\text{pos}}, R_i^{\text{neg}}\}$, explaining the analysis from the perspectives of sarcasm and non-sarcasm.

Specifically, the prompt template for facilitating debates among LLMs is: “Given you an image and text: [T]. Please provide a simplified explanation related to the text and image using contextual background knowledge, from a [Label] sentiment perspective.”. LLaVA-v1.5-13b [14] as the debater, [T] is substituted with t_i , and [Label] is designated as the specific

sentiment label y_i . A special token “ $\langle image \rangle$ ” is inserted before the prompt template to input the image I_i .

Through this prompt template, each sentiment tendency corresponds to a rational argumentative explanation, thereby providing open-world knowledge for analyzing the embedded sentiment information.

B. Multimodal Judge Module

Given an image-text pair $\langle V_i, T_i \rangle$ and the corresponding debate $R_i = \{R_i^{pos}, R_i^{neg}\}$, we first concatenate the original text and the corresponding debate as the input text of our Multimodal Judge Module as:

$$\hat{T} = [T, R^{pos}, R^{neg}] \quad (1)$$

where $[\cdot, \cdot, \cdot]$ denotes the concatenation operation.

Then we encode the input text T and the corresponding image I to get their representations as follows:

$$H_I = \text{CLIP}_{\text{vis}}(I) \quad (2)$$

$$H_{\hat{T}}^0 = \text{T5}_{\text{base}}(\hat{T}) \quad (3)$$

where $\text{CLIP}_{\text{vis}}(\cdot)$ is the visual encoder of CLIP [15] with frozen parameters. $\text{T5}_{\text{base}}(\cdot)$ denotes the text embedding layer of the LM Encoder. The output embeddings are defined as $H_I \in \mathbb{R}^{m \times d}$, $H_{\hat{T}}^0 \in \mathbb{R}^{n \times d}$, where m denotes the number of image patches, n stands for the sequence length of T , d is the mapped dimension.

With well-trained unimodal feature extractor, we stack the LLMs Debate-Guide Learning (LDG) module on top of the emotion-related feature $H_{\hat{T}}^0$, and through iterative computation, the debate effectively guides the integration of useful sarcastic features to generate the final sentiment representation. In each iteration, two interactive blocks are involved: a) Cross-Attention Module, b) Information Integrate Module. Figure 2 shows the detail process of LDG module.

To support the semantic alignment between the argument justification and the original sample for a better understanding of the cross-modal context, we utilize a cross-attention mechanism to focus on the visual representation of the text. In contrast to previous complicated fusion methods [8], [13], [16], we employ a simple one-head attention mechanism in each layer of T5 encoder, calculated as follows:

$$H_I^i = \text{softmax} \left(\frac{(H_{\hat{T}} W_Q)}{\sqrt{d}} (H_I W_K) \right) (H_I W_V) \quad (4)$$

where $W_Q \in \mathbb{R}^{d \times d}$, $W_K \in \mathbb{R}^{d \times d}$ and $W_V \in \mathbb{R}^{d \times d}$ are query, key, and value projection matrices, respectively.

The debate is then iteratively guided to obtain a representation of the interaction between H_I^i and $H_{\hat{T}}^i$:

$$H_{\hat{T}}^{i+1} = \text{Integrate}(\text{TE}^i(H_{\hat{T}}^i), H_I^i W_O) \quad (5)$$

where $\text{TE}^i(\cdot)$ is the i -th layer of the LM Encoder, H_I^i is the fixed corresponding visual feature, $i \in [0, L-1]$ and given the total L layers in the LM Encoder. $W_O \in \mathbb{R}^{d \times d}$ denotes the linear projection. We denote $H_{\hat{T}}^{i+1}$ as the final interplay representations.

TABLE I: Experimental results for sarcasm detection on in-domain dataset. We use \dagger to represent models that introduce additional knowledge by themselves, such as OCR text [9].

Model	Ref.	MMSD			
		Acc(%)	Pre(%)	Rec(%)	F1(%)
Image Modality					
Resnet [17]	CVPR'16	64.76	54.41	70.80	61.53
ViT [18]	ICLR'21	67.83	57.93	70.07	63.43
Text Modality					
TextCNN [19]	EMNLP'14	80.03	74.29	76.39	75.32
Bi-LSTM [8]	ACL'22	81.90	76.66	78.42	77.53
SIARN [20]	ACL'18	80.57	75.55	75.70	75.63
SMSD [21]	WWW'19	80.90	76.46	75.18	75.82
BERT [22]	NAACL'19	83.85	78.72	82.27	80.22
Multi-model Modality					
HFM [2]	ACL'19	83.44	76.57	84.15	80.18
D&R Net [5]†	ACL'20	84.02	77.97	83.42	80.60
Res-BERT [6]	EMNLP'20	84.80	77.80	84.15	80.85
Att-BERT [6]	EMNLP'20	86.05	78.63	83.31	80.90
InCrossMGs [7]	MM'21	86.10	81.38	84.36	82.84
CMGCN [8]	ACL'22	86.54	87.55	83.63	82.73
HKEmodel [10]†	EMNLP'22	87.36	81.84	86.48	84.09
MILNet [11]†	AAAI'23	89.50	85.16	89.16	87.11
DIP [4]	CVPR'23	89.59	87.76	86.58	87.17
DGP [9]†	IJCAI'24	87.21	87.10	86.48	86.75
DMSD-CL [12]†	AAAI'24	88.95	84.89	87.90	86.37
G*2SAM [13]	AAAI'24	90.48	87.95	89.02	88.48
Ours†	-	92.34	90.73	91.05	90.88

TABLE II: Statistics of Dataset HFM.

	Train	Val	Test
Positive	8642	959	959
Negative	11174	1451	1450
All	19816	2410	2409

C. Loss Function

In the training of the Multimodal Judge Module, binary cross-entropy loss is used.

$$\mathcal{L} = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})] \quad (6)$$

III. EXPERIMENTS

A. Datasets

The datasets involved in our experiments are described in detail below. For a fair comparison, all datasets were processed in the same manner as in previous work.

a) *HFM*: The dataset [2], one of the most popular datasets for sarcasm detection, contains 24,635 samples, each containing text content, corresponding images, and binary tags. A summary of the dataset statistics is shown in Table II.

b) *OOD*: This dataset [12] is based on HFM and is designed to assess the generalization of the model when the word distribution is different in the training and test settings. ChatGPT is used to modify the word-label association relationship, so that the distribution difference between the training set and the OOD test set can measure the effectiveness of the MSD model in mitigating the impact of spurious correlation.

B. Implementation Details

a) *Multimodal Debate Module*: As for debater, We utilize LLaVA [14] as a visual-language large model, with

TABLE III: We have re-implemented previous methods on the new trunk. The highest results for each group are highlighted in bold. ‡ denotes the result of introducing open-world knowledge generated by debate.

Modality	Models	HFM			
		Acc(%)	Pre(%)	Rec(%)	F1(%)
Image+Text (ViT + BERT)	DIP	88.03	87.64	87.56	87.60
	DIP‡	89.80	89.57	89.25	89.40
	G ² SAM	90.48	87.95	89.02	88.48
	G ² SAM‡	91.45	89.08	90.09	89.58
	Ours	88.12	89.72	86.11	87.29
	Ours‡	90.71	89.74	89.78	89.76
Image+Text (Resnet + T5)	DIP	88.49	88.06	88.16	88.11
	DIP‡	89.59	89.06	89.68	89.32
	G ² SAM	90.60	88.45	88.54	88.50
	G ² SAM‡	91.45	90.74	88.03	89.37
	Ours	89.00	88.72	87.90	88.94
	Ours‡	91.61	89.91	89.97	89.94
Image+Text (CLIP + T5)	DIP	89.97	89.37	88.73	89.01
	DIP‡	90.81	90.29	90.80	90.51
	G ² SAM	90.77	87.65	90.09	88.85
	G ² SAM‡	91.40	89.27	89.64	89.45
	Ours	89.64	88.46	87.71	88.03
	Ours‡	92.34	90.73	91.05	90.88

TABLE IV: Experimental results for MSD on OOD dataset.

Model	OOD			
	Acc(%)	Pre(%)	Rec(%)	F1(%)
DIP	71.75	70.43	69.51	69.82
DIP‡	73.75	73.32	74.35	73.31
G ² SAM	71.50	62.09	71.52	66.47
G ² SAM‡	72.75	67.07	70.13	71.01
DMSD-CL	70.25	70.41	71.34	69.96
DMSD-CL‡	73.25	74.21	75.25	73.14
Ours	73.50	72.11	72.85	72.29
Ours‡	75.50	75.96	77.11	75.32

the version “LLaVA-13b-v1.5”. To facilitate reproduction, we set the temperature to 0 and cancel the random sampling. The maximum output length is set to 256 to facilitate the processing of the judge module.

b) *Multimodal Judge Module*: Our model is implemented using PyTorch. For image encoding, we employ the pre-trained model CLIP [15] as the base model, specifically the “CLIP-ViT-B/32” version, and freeze its parameters during the training process. For text encoding, we follow [23], adopt Flan-t5 [24] as the base model, specifically the “flan-t5-base” version with 12 layers. The maximum length of text input is set to 512. During the training phase of the judge module, we using the AdamW optimizer for 15 epochs with a learning rate of 1e-4 and the drop rate is 0.1 to mitigate overfitting. The judge module was executed on a single Nvidia RTX 3090 (24G). All hyperparameters were tuned based on the validation set results.

C. Overall Result

On HFM dataset, LDGNet comprehensively surpasses the existing methods as show in Table I. To ensure a fair and comprehensive validation of the performance of these methods, two models with suboptimal performance were selected for further comparison, and four backbones were used in the

TABLE V: The ablation results of LDGNet.

Debate View		HFM		OOD	
Positive	Negative	Acc(%)	F1(%)	Acc(%)	F1(%)
✓		83.78	83.09	69.50	66.56
	✓	83.15	82.52	64.75	63.76
✓	✓	92.34	90.88	75.50	75.32

TABLE VI: The performance with different intergate methods.

Integrate	HFM		OOD	
	Acc(%)	F1(%)	Acc(%)	F1(%)
Concat	91.28	90.67	73.25	72.87
Add	92.34	90.88	75.50	75.32

experiment: ResNet [17], BERT [22], CLIP [15], and T5 [24]. Note that TableIII indicates that we have re-implemented these methods on the new trunk. In order to better evaluate the generalization performance of the model, the three suboptimal models were compared on the OOD dataset, as shown in Table IV. And when the sarcasm detection model has enough open-world knowledge, the performance is improved.

D. Ablation Study

To verify the effectiveness of the iterative LDG module, as shown in Table V, when there is only one debate view, negative or positive debate views are significantly lower on their own than at the same time. Demonstrate the effectiveness of the proposed debate model for MSD, through the multimodal debate between LLMs and the multimodal iterative integrate of small LMs. However, the model without debate achieved higher results than the model with only one debate perspective, suggesting that the guidance of only one debate perspective changed the model’s tendencies.

E. Iterative Integrate Exploring

The most critical thing in the iterative algorithm is the integration of the debate information that determines the sentiment representation, for which we have verified two simple and effective methods: Concat and Add. As shown in Table VI, Add is more effective than Concat, and Concat’s subsequent dimensionality reduction operations generate additional training parameters, so Add is finally selected.

IV. CONCLUSIONS

In this study, we investigate how sarcasm detection can be enhanced through the assistance of open-world knowledge. To this end, we first discuss the limitations associated with the current approach and propose an LDGNet model to generate effective open-world knowledge through debates between advanced LLMs to generate conflicting rationales from both sarcasm and non-sarcasm arguments. iteratively incorporating the debate process into the emotional learning process. Experimental results on HFM dataset and out-of distribution dataset with four common evaluation metrics show that our proposed model outperforms all state-of-the-art MSD models.

REFERENCES

- [1] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1136–1145. [Online]. Available: <https://doi.org/10.1145/2964284.2964321>
- [2] 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*, A. Korhonen, D. Traum, and L. Márquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2506–2515. [Online]. Available: <https://aclanthology.org/P19-1239>
- [3] R. Das and T. D. Singh, "Multimodal sentiment analysis: A survey of methods, trends, and challenges," *ACM Comput. Surv.*, vol. 55, no. 13s, jul 2023. [Online]. Available: <https://doi.org/10.1145/3586075>
- [4] C. Wen, G. Jia, and J. Yang, "Dip: Dual incongruity perceiving network for sarcasm detection," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2540–2550.
- [5] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 3777–3786. [Online]. Available: <https://aclanthology.org/2020.acl-main.349>
- [6] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1383–1392. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.124>
- [7] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 4707–4715. [Online]. Available: <https://doi.org/10.1145/3474085.3475190>
- [8] 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)*. Association for Computational Linguistics, May 2022, pp. 1767–1777. [Online]. Available: <https://aclanthology.org/2022.acl-long.124>
- [9] H. Ma, D. He, X. Wang, D. Jin, M. Ge, and L. Wang, "Multi-modal sarcasm detection based on dual generative processes," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 2279–2287, main Track. [Online]. Available: <https://doi.org/10.24963/ijcai.2024/252>
- [10] H. Liu, W. Wang, and H. Li, "Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4995–5006. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.333>
- [11] 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 Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23. AAAI Press, 2023. [Online]. Available: <https://doi.org/10.1609/aaai.v37i8.26138>
- [12] M. Jia, C. Xie, and L. Jing, "Debiasing multimodal sarcasm detection with contrastive learning," in *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada*, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 18 354–18 362.
- [13] Y. Wei, S. Yuan, H. Zhou, L. Wang, Z. Yan, R. Yang, and M. Chen, "G²sam: Graph-based global semantic awareness method for multimodal sarcasm detection," in *AAAI Conference on Artificial Intelligence*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:268692532>
- [14] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2024.
- [15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
- [16] Y. Tian, N. Xu, R. Zhang, and W. Mao, "Dynamic routing transformer network for multimodal sarcasm detection," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 2468–2480. [Online]. Available: <https://aclanthology.org/2023.acl-long.139>
- [17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [19] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: <https://aclanthology.org/D14-1181>
- [20] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1010–1020. [Online]. Available: <https://aclanthology.org/P18-1093>
- [21] 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*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2115–2124. [Online]. Available: <https://doi.org/10.1145/3308558.3313735>
- [22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [23] H. Lin, Z. Luo, W. Gao, J. Ma, B. Wang, and R. Yang, "Towards explainable harmful meme detection through multimodal debate between large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2401.13298>
- [24] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, D. Valter, S. Narang, G. Mishra, A. W. Yu, V. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Hsin Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *ArXiv*, vol. abs/2210.11416, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:253018554>