Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 2, 2025

Multimodal Affective Computing Method with Cross-Modal Attention

Bo He[a], Hao Han[*], Hongqian Zhang[b], Weining Bai[c]

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 40054, China

[*] **Corresponding author**: Hao Han (Email: hanhao0731@outlook.com)

[a]hebo@cqut.edu.cn, [b]1329143810@outlook.com, [c]bwn52240313232@163.com

Abstract: Multimodal affective computing (MAC) aims to enable machines to recognize and understand human emotions by integrating information from multiple modalities, including text, speech, facial expressions, and gestures. Traditional fusion methods such as early fusion, late fusion, and tensor fusion often struggle to capture fine-grained inter-modal dependencies and handle conflicts between modalities. Cross-modal attention has emerged as a powerful mechanism to address these challenges by selectively aligning and weighting features across modalities, highlighting relevant cues while suppressing noise. This survey reviews recent advances in MAC with a focus on cross-modal attention, covering core tasks such as Multimodal Sentiment Analysis (MSA), Multimodal Emotion Recognition in Conversations (MERC), and Multimodal Sarcasm and Humor Detection (MSD). We analyze the principles of cross-modal attention, summarize representative models, and discuss their empirical performance. Finally, we identify current challenges, including dataset limitations, asynchronous modality alignment, and interpretability, and propose future directions such as large-scale pretraining, knowledge-enhanced modeling, and interpretable attention mechanisms. Overall, cross-modal attention significantly improves robustness, context-awareness, and fine-grained emotion understanding in multimodal affective computing.

Keywords: Multimodal Affective Computing, Cross-Modal Attention, Sentiment Analysis, Emotion Recognition in Conversation, Sarcasm Detection

1. Introduction

Affective computing, which aims to endow machines with the ability to recognize and respond to human emotions, plays a pivotal role in applications such as social media analysis, dialogue systems, and human–computer interaction. Human affect is inherently multimodal, expressed through text, speech, facial expressions, and other behavioral cues. Relying on a single modality often fails to capture the richness and ambiguity of emotional expressions, leading to reduced robustness and generalization. To address these limitations, multimodal affective computing (MAC) has emerged as a promising direction. By integrating information from different modalities, MAC achieves more reliable emotion understanding compared with unimodal approaches. Traditional fusion strategies, including early fusion, late fusion, and tensor fusion, have advanced the field, yet they struggle to capture fine-grained cross-modal dependencies and to resolve conflicts between modalities. These challenges call for more flexible and adaptive mechanisms for multimodal interaction.

Recently, cross-modal attention mechanisms have gained prominence as an effective solution. By selectively aligning and integrating information across modalities, cross-modal attention highlights relevant signals while suppressing noise and inconsistencies. This capability is particularly valuable for affective tasks where subtle interdependencies or modality mismatches are crucial—for example, sarcasm often emerges from incongruent verbal and visual cues.

This survey provides an overview of multimodal affective computing with a focus on cross-modal attention. First, we review the foundations of MAC and its key tasks such as sentiment and emotion recognition, dialogue emotion recognition, and sarcasm or humor detection. Second, we introduce cross-modal attention architectures and their advantages over traditional fusion methods. Third, we analyze applications of cross-modal attention in various affective computing scenarios. Finally, we discuss challenges and future directions.

2. Multimodal Affective Computing: Background

Multimodal affective computing (MAC) aims to understand and interpret human emotions by integrating information from multiple modalities, including text, speech, facial expressions, gestures, and physiological signals[1], [2]. Unlike unimodal approaches, which rely solely on a single source of information, multimodal methods leverage complementary cues across different channels, thereby improving robustness and enabling richer emotional understanding. For instance, a sarcastic comment in text may be accompanied by a tone of voice or facial expression that conveys the opposite sentiment, which cannot be captured effectively by analyzing text alone[3].

2.1. Core Tasks

MAC encompasses several key tasks, reflecting the diversity of emotional expressions across different modalities. In this paper, we focus on introducing the following task types:

Multimodal Sentiment Analysis (MSA) focuses on classifying the sentiment or emotional state expressed in videos, social media posts, or short texts by combining textual, auditory, and visual features[4]. Typical applications include movie review analysis, online comment moderation, and affect-aware recommendation systems.

Multimodal Emotion Recognition in Conversations (MERC) aims to model emotion dynamics in multi-turn

conversations. Here, the challenge lies not only in fusing multiple modalities, but also in capturing contextual dependencies and speaker-specific patterns over time[5]–[7]. Accurate MERC is critical for conversational AI, empathetic dialogue systems, and virtual assistants.

Multimodal Sarcasm Detection (MSD) addresses scenarios where emotional expression is implied through incongruence between modalities, such as a positive textual comment accompanied by a negative facial expression or image[3], [8], [9]. Detecting such subtle inconsistencies requires models that can align and compare information across modalities effectively.

2.2. Traditional Fusion Approaches

Early studies in MAC relied on conventional fusion strategies to integrate information from different modalities. Early fusion concatenates raw or preprocessed features from all modalities before feeding them into a classifier[10]–[12]. While straightforward, this approach often suffers from noise propagation and imbalance between modalities. Late fusion processes each modality independently and combines their predictions at the decision level[13], [14], offering robustness but limited capacity to model inter-modal interactions. To capture higher-order correlations, tensor fusion networks (TFN) and other hybrid fusion methods were proposed[15], [16], allowing multiplicative interactions among modalities. However, these approaches tend to be computationally expensive and still struggle with modeling dynamic dependencies and fine-grained interactions.

2.3. Limitations and Motivation for Cross-Modal Modeling

Despite these advances, traditional multimodal fusion methods face several challenges. First, they often fail to capture complex interdependencies between modalities, especially when emotions are expressed through subtle or conflicting cues. Second, they provide limited interpretability, making it difficult to understand how each modality contributes to the final prediction. Third, handling temporal context and long sequences remains problematic, particularly in dialogue-based or video-based emotion recognition. These limitations motivate the development of more sophisticated mechanisms for cross-modal interaction, with attention-based methods emerging as a promising solution for effectively aligning and weighting information across modalities.

3. Cross-modal Attention: Foundations

3.1. A Brief Review of Attention Mechanisms

The attention mechanism was originally introduced in natural language processing to improve sequence modeling by allowing models to focus on the most relevant parts of the input[17]. Self-attention, popularized by the Transformer architecture, enables a sequence element to attend to all other elements within the same modality, capturing global dependencies and contextual information. This has proven highly effective in a range of tasks including machine translation[18], [19], text classification[20], [21], and text summarization[22], [23].

In multimodal learning, however, modeling intra-modal relationships alone is insufficient. To address this, co-attention mechanisms were proposed, where two different modalities attend to each other in a coordinated manner[8], [13], [17], [24]. Co-attention allows one modality (e.g., text) to guide the focus over another modality (e.g., visual regions in an image), and vice versa, enabling richer cross-modal interactions. These ideas laid the foundation for more advanced cross-modal attention methods.

3.2. Principles of Cross-modal Attention

Cross-modal attention refers to the explicit modeling of information exchange across different modalities through selective alignment and feature weighting. Instead of naïvely concatenating or averaging features, cross-modal attention dynamically learns which features from one modality should be highlighted given the context of another.

Formally, given a query representation $Q \in \mathbb{R}^{n\times d}$ from one modality (e.g., text) and key–value pairs $(K,V) \in \mathbb{R}^{m\times d}$ from another modality (e.g., visual or acoustic signals), the cross-modal attention can be defined as:

$$Att(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

where the similarity between query and key determines the attention weights, and the values provide the contextualized features.

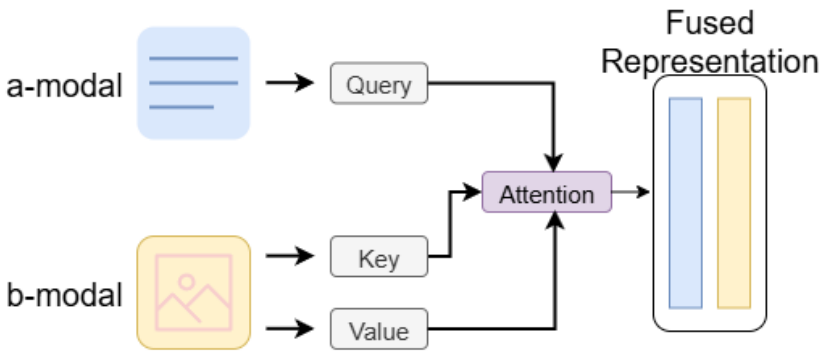

Fig 1. Cross-Modal Attention Schematic

In practice, bi-directional cross-modal attention can be established by performing this operation in both directions (as illustrated in Figure 1):

$$H^{(T)} = Att\left(Q^{(T)}, K^{(V)}, V^{(V)}\right) \quad (2)$$

$$H^{(V)} = Att\left(Q^{(V)}, K^{(T)}, V^{(T)}\right) \quad (3)$$

where T and V denote text and visual modalities, respectively. This allows each modality to refine its representation conditioned on the other, yielding more semantically aligned features.

The key principle is information enhancement through selective alignment: each modality not only contributes its own features but also shapes the representation of other

modalities by emphasizing complementary or contradictory signals. For affective computing, this is crucial since emotions are often conveyed through subtle interdependencies[25]–[27] (e.g., smiles accompanying positive text) or inconsistencies[28], [29] (e.g., sarcasm expressed by mismatch between words and tone).

3.3. Advantages over Traditional Fusion

Compared with early, late, or tensor fusion methods, cross-modal attention provides several advantages:

- Highlighting inter-modal relevance: It selectively emphasizes features that are consistent or complementary across modalities.
- Revealing conflicts: It enables the model to exploit modality mismatches, which are important signals in sarcasm or irony detection.
- Leveraging complementarity: It naturally integrates multimodal information at multiple levels of granularity, reducing redundancy and noise.

Overall, cross-modal attention has become a cornerstone technique in multimodal affective computing, enabling models to move beyond shallow fusion and toward more context-aware and interpretable emotion understanding.

4. Applications in Multimodal Affective Computing

Cross-modal attention mechanisms have been extensively applied across various task types within multimodal affective computing. The following subsections shall respectively examine their typical applications and value in Multimodal Sentiment Analysis (MSA), Multimodal Emotion Recognition in Conversations (MERC), and Multimodal Sarcasm Detection (MSD).

4.1. Multimodal Sentiment Analysis (MSA)

Multimodal Sentiment Analysis (MSA) focuses on classifying the sentiment or emotional state expressed through videos, social media posts, or short texts by leveraging multiple modalities, such as textual content, speech, and visual signals. Compared with unimodal approaches, MSA benefits from integrating complementary cues across modalities, leading to more robust and nuanced sentiment predictions.

Challenges: MSA faces several challenges due to the heterogeneous nature of modalities. First, feature alignment across text, audio, and visual streams is non-trivial, as different modalities often operate at distinct temporal resolutions. Second, modality noise or irrelevant signals can interfere with accurate sentiment classification. Third, the relative importance of each modality may vary across instances, necessitating dynamic weighting.

Role of Cross-modal Attention: Cross-modal attention has been widely adopted in MSA to address these challenges. Its primary function is to align features across modalities and dynamically assign weights to each modality, highlighting relevant signals while suppressing noise. For example,

- Ghosal et al. proposed the Multimodal Multi-Utterance Bi-modal Attention (MMUU-BA) framework[30], which leverages both self-attention and cross-modal attention mechanisms to capture latent emotions within sentences.
- Curto et al. designed Dyadformer, inspired by Transformer encoders[31], which employs cross-attention for modality fusion and facilitates information flow between interlocutors.
- Zhuang et al. introduced GLoMo[32], which first integrates global and local representations via a Mixture-of-Experts (MoE) mechanism and then fuses feature information using cross-modal attention.
- Zhu et al. proposed KEBR[33], incorporating a text-dominant cross-modal attention module to enhance multimodal feature representations.

Applications and Empirical Evidence: Cross-modal attention models have achieved superior performance on benchmark datasets, outperforming traditional early and late fusion methods. Dynamic modality weighting improves the model's ability to handle noisy or missing inputs, while fine-grained alignment enhances recognition of nuanced sentiment expressions.

In summary, the key contributions of cross-modal attention in MSA are:

- Feature alignment – ensuring correspondence across heterogeneous modalities.
- Dynamic modality weighting – emphasizing informative modalities while reducing noise.
- Fine-grained emotional cue extraction – capturing subtle expressions across text, audio, and visual channels.

These mechanisms allow MSA models to robustly and accurately predict sentiment in complex multimodal scenarios.

4.2. Multimodal Emotion Recognition in Conversations (MERC)

Multimodal Emotion Recognition in Conversations (MERC) extends the task of multimodal sentiment analysis to a more challenging and dynamic setting: multi-turn dialogues. Unlike single-turn utterances, conversations involve evolving emotional dynamics, speaker-specific patterns, and dependencies across time. These characteristics make MERC a crucial task for applications such as empathetic conversational agents, affect-aware virtual assistants, and social robots. Cross-modal attention has become a central component in MERC because it can explicitly model temporal alignment, contextual dependencies, and speaker-aware interactions across modalities.

Challenges: One of the main difficulties in MERC lies in capturing long-range contextual information, since the emotional state of a given utterance often depends on several previous turns. Another challenge is speaker dependency: different speakers may display distinct emotional trajectories, and interactions between them shape the overall dialogue sentiment. Finally, ensuring temporal consistency is crucial, as abrupt or unrealistic emotion shifts can negatively affect recognition accuracy.

Role of Cross-modal Attention: Cross-modal attention has been widely adopted in MERC to address these challenges. Its main advantage is the ability to align and integrate multimodal features across time while selectively emphasizing the most relevant signals. For example,

- Zhang et al. proposed CMCF-SRNet[27], a dialogue modeling framework that leverages attention mechanisms and semantic graphs. It employs a semantic-graph-based Transformer encoder to effectively capture latent emotions.
- Chudasama et al. introduced M2FNet[14], which designs a feature extractor based on multi-head

attention to capture latent features across different modalities.

- Shi et al. proposed the MultiEMO framework[34], which integrates multimodal cues effectively by modeling cross-domain relationships through bidirectional multi-head cross-attention layers.
- Yun et al. developed TelME[35], which utilizes knowledge distillation to transfer knowledge from a text-modality teacher model to non-text modality student models, thereby enhancing the performance of weaker modalities. Finally, it employs attention-based multimodal shifted fusion to integrate features from multiple modalities.

Applications and Empirical Evidence: On benchmark datasets such as IEMOCAP and MELD, cross-modal attention-based models consistently outperform early fusion and late fusion approaches. They have proven especially effective for complex emotions like frustration, anger, or empathy, which are difficult to capture from a single modality alone. Furthermore, hierarchical and speaker-aware cross-modal attention strategies have shown clear benefits in modeling realistic conversational scenarios.

In summary, cross-modal attention in MERC plays a key role in:

- Temporal alignment – linking current utterances with historical multimodal cues.
- Dialogue-level emotion tracking – modeling evolving dynamics over multiple turns.
- Speaker-aware modeling – capturing individual-specific and interpersonal emotional dependencies.

These advances illustrate how cross-modal attention can effectively bridge the gap between multimodal feature fusion and dialogue-level emotional reasoning.

4.3. Multimodal Sarcasm Detection (MSD)

Multimodal Sarcasm and Humor Detection (MSD) aims to identify sarcastic or humorous expressions that are often conveyed through incongruences between modalities. For example, a positive textual comment may be accompanied by a negative facial expression or a contrasting tone of voice. Such subtle modality conflicts make MSD particularly challenging and highly relevant for social media analysis and meme understanding.

Challenges: MSD presents several unique challenges. First, detecting sarcasm or humor relies on recognizing modality conflicts, which may be implicit or context-dependent. Second, semantic subtleties in text or speech often require alignment with visual cues to uncover the underlying affect. Third, datasets for MSD are limited and highly variable, complicating model generalization.

Role of Cross-modal Attention: Cross-modal attention has proven effective in addressing these challenges by explicitly modeling interactions and inconsistencies between modalities. For instance,

- Pan et al. developed a BERT-based model[29] that captures contextual information in the task by considering both intra-modal and inter-modal inconsistencies.
- Xu et al. introduced the Decomposition and Relation (D & R) network[9], which models both shared and modality-specific features in visual-text bimodal data. By incorporating attention mechanisms, this approach effectively captures visual and textual features, enabling more comprehensive multimodal sarcasm detection.
- Desai et al. proposed a Transformer-based encoder[36] that leverages cross-modal attention to focus on distinctive features across different modalities.
- Ou et al. presented a global–local cross-modal[37] interaction learning method based on CLIP, which captures multi-granularity cues and enhances cross-modal interactions in multimodal sarcasm detection tasks.

Applications and Empirical Evidence: Cross-modal attention enables MSD models to focus on the most informative features that indicate incongruence, such as a smiling face paired with sarcastic text. Models utilizing this mechanism have achieved higher accuracy and better robustness on multimodal sarcasm and humor datasets. These approaches also offer interpretability, as attention weights can indicate which modality or feature contributes to the detection of sarcasm or humor.

In summary, cross-modal attention contributes to MSD by:

- Detecting modality conflicts – uncovering discrepancies between textual, visual, and auditory cues.
- Capturing implicit semantics – aligning subtle cues across modalities to reveal underlying sarcasm or humor.
- Improving interpretability – providing insights into which features are decisive for detection.

5. Conclusion and Future Directions

Multimodal affective computing has witnessed significant advancements in recent years, with cross-modal attention mechanisms emerging as a central technique for effectively integrating heterogeneous modalities. By selectively aligning and weighting information from text, speech, and visual channels, cross-modal attention enables models to capture subtle emotional cues, resolve conflicts across modalities, and model context-dependent or temporal dynamics. Its application spans key tasks including Multimodal Sentiment Analysis (MSA), Multimodal Emotion Recognition in Conversations (MERC), and Multimodal Sarcasm and Humor Detection (MSD), demonstrating consistent performance improvements over traditional early or late fusion approaches. Overall, cross-modal attention has proven to be invaluable for robust, interpretable, and fine-grained emotion understanding.

Despite these advances, several challenges remain. First, dataset limitations constrain the generalization and scalability of multimodal models, especially in sarcasm and humor detection or low-resource domains. Second, cross-modal alignment remains difficult due to asynchronous modalities, heterogeneous feature spaces, and temporal mismatches. Third, while attention weights provide some interpretability, the overall explainability of cross-modal interactions is still limited, hindering insight into how models make affective predictions.

Looking forward, several promising directions can further advance the field:

- Large-scale pretraining and multimodal foundation models: Leveraging pretrained models across modalities can improve generalization, enable zero-shot or few-shot learning, and reduce reliance on task-specific annotated datasets.
- Knowledge-enhanced multimodal modeling: Incorporating external knowledge sources, such as knowledge graphs or commonsense reasoning, can

enhance semantic understanding and improve recognition of complex affective phenomena like sarcasm and humor.

- Interpretable and controllable attention mechanisms: Developing attention architectures that provide transparent explanations of cross-modal interactions can increase trustworthiness and facilitate debugging in practical applications.
- Generalization and low-resource scenarios: Techniques for domain adaptation, transfer learning, and semi-supervised learning can help cross-modal attention models perform effectively in unseen or data-scarce environments.

In conclusion, cross-modal attention has fundamentally transformed multimodal affective computing, enabling more accurate, context-aware, and nuanced emotion recognition. By addressing current challenges and exploring the future directions outlined above, the field is poised to achieve even greater advances, with broad implications for human–computer interaction, social media understanding, and affect-aware AI systems.

Acknowledgements

This work is funded by the humanities and social sciences research project of the ministry of education (No.24YJA870003) and the Chongqing University of Technology Graduate Education High Quality Development Project (No. gzlcx20253209)

References

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, Sep. 2017.

[2] S. K. D'mello and J. Kory, "A Review and Meta-Analysis of Multimodal Affect Detection Systems," ACM Comput. Surv., vol. 47, no. 3, pp. 1–36, Apr. 2015.

[3] 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, Amsterdam The Netherlands, 2016, pp. 1136–1145.

[4] R. Das and T. D. Singh, "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges," ACM Comput. Surv., vol. 55, no. 13s, pp. 1–38, Dec. 2023.

[5] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," presented at the Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers), 2017, pp. 873–883.

[6] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," presented at the Proceedings of the AAAI conference on artificial intelligence, 2019, vol. 33, pp. 6818–6825.

[7] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," presented at the Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, 2018, vol. 2018, p. 2122.

[8] 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, Florence, Italy, 2019.

[9] 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, Online, 2020.

[10] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 2021, pp. 5666–5675.

[11] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring Emotion from Conversational Voice Data: A Semi-Supervised Multi-Path Generative Neural Network Approach," AAAI, vol. 32, no. 1, Apr. 2018.

[12] J. Li, X. Wang, Y. Liu, and Z. Zeng, "CFN-ESA: A Cross-Modal Fusion Network With Emotion-Shift Awareness for Dialogue Emotion Recognition," IEEE Trans. Affective Comput., vol. 15, no. 4, pp. 1919–1933, Oct. 2024.

[13] S. Dutta and S. Ganapathy, "HCAM -- Hierarchical Cross Attention Model for Multi-modal Emotion Recognition." arXiv, 09-Jan-2024.

[14] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 4651–4660.

[15] T. M. Tashu, S. Hajiyeva, and T. Horvath, "Multimodal Emotion Recognition from Art Using Sequential Co-Attention," J. Imaging, vol. 7, no. 8, p. 157, Aug. 2021.

[16] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," Knowledge-Based Systems, vol. 167, pp. 26–37, Mar. 2019.

[17] A. Galassi, M. Lippi, and P. Torroni, "Attention in Natural Language Processing," IEEE Trans. Neural Netw. Learning Syst., vol. 32, no. 10, pp. 4291–4308, Oct. 2021.

[18] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 5797–5808.

[19] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What Does BERT Look at? An Analysis of BERT's Attention," in Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence, Italy, 2019, pp. 276–286.

[20] S. Serrano and N. A. Smith, "Is Attention Interpretable?," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, 2016, pp. 1480–1489.

[22] A. Lauscher, G. Glavaš, S. P. Ponzetto, and K. Eckert, "Investigating the Role of Argumentation in the Rhetorical Analysis of Scientific Publications with Neural Multi-Task Learning Models," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, pp. 3326–3338.

[23] L. Wu, F. Tian, L. Zhao, J. Lai, and T.-Y. Liu, “Word Attention for Sequence to Sequence Text Understanding,” AAAI, vol. 32, no. 1, Apr. 2018.

[24] J. Cheng, I. Fostiropoulos, B. Boehm, and M. Soleymani, “Multimodal Phased Transformer for Sentiment Analysis,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021, pp. 2447–2458.

[25] C. Liu, H. Ding, Y. Zhang, and X. Jiang, “Multi-Modal Mutual Attention and Iterative Interaction for Referring Image Segmentation,” IEEE Trans. on Image Process., vol. 32, pp. 3054–3065, 2023.

[26] Y. Jing and X. Zhao, “DQ-Former: Querying Transformer with Dynamic Modality Priority for Cognitive-aligned Multimodal Emotion Recognition in Conversation,” in Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne VIC Australia, 2024, pp. 4795–4804.

[27] X. Zhang and Y. Li, “A Cross-Modality Context Fusion and Semantic Refinement Network for Emotion Recognition in Conversation,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 13099–13110.

[28] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, and R. Xu, “MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System,” in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, 2023, pp. 10834–10845.

[29] 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, Online, 2020.

[30] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, “Contextual Inter-modal Attention for Multi-modal Sentiment Analysis,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, pp. 3454–3466.

[31] D. Curto, A. Clapes, J. Selva, S. Smeureanu, J. C. S. Jacques Junior, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund, S. Escalera, and C. Palmero, “Dyadformer: A Multi-modal Transformer for Long-Range Modeling of Dyadic Interactions,” in 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 2177–2188.

[32] Y. Zhuang, Y. Zhang, Z. Hu, X. Zhang, J. Deng, and F. Ren, “GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis,” in Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne VIC Australia, 2024, pp. 1800–1809.

[33] A. Zhu, M. Hu, X. Wang, J. Yang, Y. Tang, and F. Ren, “KEBR: Knowledge Enhanced Self-Supervised Balanced Representation for Multimodal Sentiment Analysis,” in Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne VIC Australia, 2024, pp. 5732–5741.

[34] T. Shi and S.-L. Huang, “MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023, pp. 14752–14766.

[35] T. Yun, H. Lim, J. Lee, and M. Song, “TelME: Teacher-leading Multimodal Fusion Network for Emotion Recognition in Conversation,” in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 2024, pp. 82–95.

[36] P. Desai, T. Chakraborty, and M. S. Akhtar, “Nice Perfume. How Long Did You Marinate in It? Multimodal Sarcasm Explanation,” AAAI, vol. 36, no. 10, pp. 10563–10571, Jun. 2022.

[37] L. Ou and Z. Li, “Modeling Multi-Task Joint Training of Aggregate Networks for Multi-Modal Sarcasm Detection,” in Proceedings of the 2024 International Conference on Multimedia Retrieval, Phuket Thailand, 2024, pp. 833–841.