Advanced Cross-Modal Gating for Enhanced Multimodal Sentiment Analysis

Emily Johnson , Rodolfo Patel , Michael Smith *

Posted Date: 6 August 2024

doi: 10.20944/preprints202408.0265.v1

Keywords: multimodal sentiment analysis; cross-modal gating; deep learning fusion

 doi:10.20944/preprints202408.0265.v1

Article

Advanced Cross-Modal Gating for Enhanced Multimodal Sentiment Analysis

Emily Johnson, Rodolfo Patel and Michael Smith *

Briar Cliff University

* Correspondence: michael.smith@briarcliff.edu

Abstract: The rapidly evolving domain of multimodal sentiment analysis is crucial for unraveling the intricate layers of emotional expression in social media content, customer service interactions, and personal vlogs. This research introduces a cutting-edge Advanced Cross-Modal Gating (ACMG) framework that significantly enhances the precision of sentiment analysis by refining the interplay among textual, auditory, and visual modalities. Our approach addresses three foundational aspects of sentiment analysis: (1) Advanced learning of cross-modal interactions, which focuses on extracting and synthesizing sentiment from varied modal inputs, thus providing a holistic view of expressed emotions; (2) Mastery over the temporal dynamics of multimodal data, enabling the model to maintain context and sentiment continuity over extended interactions; and (3) Deployment of a novel fusion strategy that not only integrates unimodal and cross-modal cues but also dynamically adjusts the influence of each modality based on its contextual relevance to the sentiment being expressed. The exploration of these dimensions reveals that the nuanced modeling of cross-modal interactions is crucial for enhancing model responsiveness and accuracy. By applying the ACMG model to two highly regarded datasets—CMU Multimodal Opinion Level Sentiment Intensity (CMU-MOSI) and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI)—we achieve groundbreaking accuracies of 83.9% and 81.1%, respectively. These results represent significant improvements of 1.6% and 1.34% over the current state-of-the-art, showcasing the superior performance and potential of our approach in navigating the complexities of multimodal sentiment analysis.

Keywords: multimodal sentiment analysis; cross-modal gating; deep learning fusion

1. Introduction

Sentiment analysis [1,2], a subfield of natural language processing (NLP), has grown tremendously in its scope and applicability, especially in understanding human emotions conveyed through language. This computational technique is designed to automatically analyze and classify the emotional tone behind words used in text data, enabling machines to understand the sentiments expressed by humans. Originally focusing primarily on textual data, sentiment analysis has expanded to encompass a wider array of sources such as social media posts, customer reviews, and news articles, where it serves crucial roles in market analysis [3], public relations, and automated customer service. The ability to parse complex human emotions automatically offers significant advantages in various industries by helping businesses gauge public sentiment, tailor marketing strategies, and improve customer experiences. As sentiment analysis evolves, its integration with other technological advances like machine learning and deep learning has markedly improved its accuracy and the granularity of sentiment detection, pushing the boundaries of how machines understand human emotions [9].

Sentiment analysis [10,11] has emerged as a pivotal area in the realm of spoken language understanding, aimed at discerning the sentiments expressed by individuals towards various subjects, such as products, events, or topics. The advent of social media platforms—including Facebook, WhatsApp, Instagram, and YouTube—has catalyzed an exponential increase in the generation of multimedia content such as podcasts, vlogs, interviews, and commentaries. This multimedia content is rich with parallel acoustic signals (e.g., vocal expressions like intensity and pitch) and visual cues (e.g., facial expressions and gestures), alongside textual information (spoken words), offering a comprehensive scaffold for advanced sentiment analysis.

 doi:10.20944/preprints202408.0265.v1

On the other hand, multimodal learning is an advanced area of machine learning that seeks to process and relate information from multiple sensory modalities—such as text, image, and sound—to better understand the context and insights that might not be accessible through unimodal data. In the digital age, where data comes in various forms, the ability to integrate and interpret this heterogeneous data is crucial. This field addresses the challenge of capturing the complementary and redundant properties of modalities that naturally occur in human communication. For instance, in human interactions, spoken words (audio), facial expressions (video), and text (transcriptions) work together to convey complete messages [17,18]. Multimodal learning models strive to create representations that can effectively merge these different types of data to perform tasks such as sentiment analysis, identity verification, and multimedia content recommendation more accurately. By harnessing the strengths of each modality, these models achieve superior performance in complex environments where unimodal signals might be insufficient or misleading, enhancing both the robustness and the depth of analytical applications.

This paper introduces the Advanced Cross-Modal Gating (ACMG) framework, designed to enhance sentiment analysis by intricately fusing multimodal data streams. Our ACMG system focuses on three strategic areas: 1) advanced learning techniques for cross-modal interactions to robustly capture and synthesize sentiments from diverse modal inputs; 2) sophisticated handling of the persistent multimodal dependencies that emerge in extended discourses; and 3) an innovative fusion strategy that not only assimilates unimodal and intermodal cues but also adapts dynamically to the contextual significance of each modality in real-time sentiment analysis.

Prior methodologies in multimodal sentiment analysis generally fall into three categories: (i) Approaches that analyze modalities independently and later fuse their outputs [19,20], (ii) Strategies that jointly analyze the interactions among two or more modalities [21,22], and (iii) Techniques that leverage both unimodal and cross-modal contributions, often employing attention mechanisms to refine this integration [27–34].

Traditionally, fusion methods like early or decision-level fusion dominated the field [19]. However, more recent studies suggest integrating fusion at various levels and hierarchies to capture the nuanced dynamics of multimodal interactions more effectively [21]. For example, multi-kernel learning has been employed to fuse acoustic and visual features with textual data, enhancing the richness of the sentiment analysis [35]. Other advanced techniques include gated multimodal embeddings with temporal attention for word-level fusion [27] and hierarchical attention architectures that build upon aligned multimodal features.

Our ACMG model innovatively builds upon these foundations by introducing a conditional gating mechanism that modulates cross-modal interactions based on linguistic content, vocal tone, and visual expressions. This mechanism selectively emphasizes the modalities most relevant to the sentiment being expressed, thereby enhancing the accuracy and contextuality of the analysis. Additionally, we incorporate a self-attention layer to capture long-term dependencies across utterances within videos, enabling unrestricted information flow and deeper contextual understanding. The fusion of these self-attended and gated cross-interaction representations through a recurrent layer results in robust, deep multimodal contextual feature vectors for each utterance.

The primary contributions of this paper are threefold: **1)** the development of a learnable gating mechanism that strategically controls information flow during cross-modal interactions; **2)** the application of self-attended contextual representations to capture extended dependencies; and **3)** a sophisticated recurrent layer-based approach for integrating self and gated cross fusion feature vectors to derive deep, modality-specific multimodal feature vectors.

2. Related Work

Sentiment analysis has evolved significantly from its initial applications, which were primarily focused on analyzing textual data such as product reviews and social media posts. Early methods relied on simple lexicon-based approaches that scored words based on predefined sentiment dic-

 doi:10.20944/preprints202408.0265.v1

tionaries [39–44]. However, these methods were limited by their inability to understand context or the subtleties of language such as irony and sarcasm. With the advancement of machine learning techniques, particularly supervised learning, sentiment analysis began to employ more sophisticated models like support vector machines (SVMs) and naive Bayes classifiers, which offered improved accuracy by learning from large datasets of labeled examples.

The integration of deep learning into sentiment analysis marked a significant leap forward, enabling the analysis of complex sentence structures and semantic nuances [48,49]. Neural networks, particularly recurrent neural networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks, have been pivotal in capturing the temporal dynamics of language. These models excel in tasks that involve understanding context over longer stretches of text, making them ideal for sentiment analysis in conversations and narratives. More recently, the emergence of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), has revolutionized sentiment analysis by providing even deeper contextual analysis capabilities through mechanisms like self-attention, which allows the model to weigh the importance of each word in a sentence relative to others.

Multimodal learning extends the capabilities of traditional machine learning by incorporating multiple types of data inputs. The foundational concept behind multimodal learning is to leverage the inherent strengths of each data modality to improve the accuracy and robustness of predictive models. Early research in the field focused on simple concatenation techniques for combining features from different modalities [55,56]. However, these early attempts often failed to capture the complex interdependencies between modalities. The realization that different modalities could provide complementary information about the same phenomenon spurred developments in more sophisticated integration techniques, such as joint feature learning and co-training methods [59–61].

Recent advances in multimodal learning [66] have focused on more dynamic interaction models that can effectively synchronize and integrate data from diverse sources. Techniques such as cross-modal attention mechanisms have been developed to selectively focus on the most relevant features across modalities, enhancing tasks like video analysis where audio and visual cues need to be synchronized [68,69]. Another significant development is the use of hybrid models that combine convolutional neural networks (CNNs) for image processing with LSTMs for sequential data like text and audio, enabling these models to handle complex multimodal inputs effectively. These advanced models are not only more adept at handling data alignment and temporal synchronization but also significantly improve performance in applications such as multimedia event detection and multimodal sentiment analysis.

In the specific domain of multimodal sentiment analysis, the integration of textual, audio, and visual data has led to more accurate detection of sentiments, as it mirrors human communication more closely. Recent studies have explored various architectures for effective multimodal integration. For instance, some approaches focus on modality fusion at different levels—feature-level, decision-level, and hybrid—each offering distinct advantages depending on the application. Other innovative approaches have included the development of end-to-end trainable systems that use multimodal deep learning to simultaneously learn feature representations and sentiment classification. These systems often employ complex gating mechanisms to manage the flow of information from different modalities, ensuring that the model remains sensitive to the most informative cues at any given moment.

3. Methodology

Our Advanced Cross-Modal Gating (ACMG) model is designed to intricately learn the interactions between different modalities, governed by a sophisticated learnable gating mechanism. The comprehensive architecture of our system has key components: *contextual utterance representation*, *self-attention*, *cross-attention*, *gating mechanism for cross-interaction*, and *deep multimodal fusion*.

 doi:10.20944/preprints202408.0265.v1

3.1. Contextual Utterance Representation

The foundation of our model starts with the extraction of rich, contextual representations from utterances within each modality. This is achieved through the application of Bi-directional Gated Recurrent Units (Bi-GRUs) [70], which are adept at capturing both past and future context. For each modality, utterance-level features are sequentially fed into a separate Bi-GRU, resulting in dynamic, modality-specific contextual representations denoted as H. Formally, the contextual utterance representations ($H_T \in R^{u \times d}$) for a sequence of utterances ($U_1, U_2, ..., U_u$) in the *Text* modality are computed as:

$$H_T = Bi\text{-}GRU(U_1, U_2, ..., U_u) \tag{1}$$

where subscript T denotes *Text*, with A and V representing *Audio* and *Video* modalities respectively. Each modality's representations capture the nuanced, time-dependent characteristics pertinent to that specific modality.

3.2. Self Attention

To address the challenge of capturing long-term dependencies, especially in videos containing up to 100 utterances, we employ a bilinear attention mechanism [71] based on self-matching layers applied to the contextual utterance representations. For *Text*, the self-attention mechanism is modeled as:

$$M_T = H_T W H_T^T,\ M_T \in R^{u \times u} \tag{2a}$$

$$A_T(i,) = softmax(M_{T_i,}) \tag{2b}$$

$$S_T = A_T.H_T,\ S_T \in R^{u \times d} \tag{2c}$$

Here, Equation 2a computes the self-matching matrix with W being a trainable matrix. Equation 2b calculates the self-attention scores for each utterance, U_i, and Equation 2c produces the self-attended utterance representations. The self-matching matrix M_T is computed through a trainable weight matrix W, with self-attention scores A_T then applied to generate self-attended representations S_T, enhancing the representation with contextual awareness of the entire sequence.

3.3. Cross Attention

Leveraging multimodal data provides a unique opportunity to learn intricate interactions between modalities. Following methods similar to those discussed by Ghosal et al. [34], our model learns cross-interaction vectors. For Text (H_T) and Video (H_V) modalities, the co-attention matrix ($M_{TV} \in R^{u \times u}$) is defined as:

$$M_{TV} = H_T W H_V^T;\ W \in R^{d \times d} \tag{3}$$

Cross-attention representations for Text ($C_{VT} \in R^{u \times d}$) and Video ($C_{TV} \in R^{u \times d}$) are subsequently computed as:

$$A_{TV}(i:) = softmax(M_{TV_{i:}}) \tag{4a}$$

$$A_{VT}(:j) = softmax(M_{TV_{:j}}) \tag{4b}$$

$$C_{VT} = A_{VT}.H_T,\ C_{TV} = A_{TV}.H_V \tag{4c}$$

3.4. Gating Mechanism for Cross Interaction

The integration of cross-modal data introduces challenges due to the potential imperfections in individual modalities. To address this, we implement a gating mechanism that selectively learns which

 doi:10.20944/preprints202408.0265.v1

cross-interactions to emphasize [72,73]. The gated cross-fused vector ($F_{PQ} \in R^{u \times d}$) for Text and Video modalities is modeled as:

$$F_{VT} = fusion(C_{VT}, H_T) \tag{5a}$$
$$F_{TV} = fusion(C_{TV}, H_V) \tag{5b}$$

The fusion function combines the cross-interaction and contextual representations using a gated mechanism:

$$X(P,Q) = tanh(W_F.[P,Q,P-Q,P \circ Q] + b_F) \tag{6a}$$
$$G(P,Q) = \sigma(W_G^T.[P,Q,P-Q,P \circ Q] + b_G) \tag{6b}$$
$$F_{PQ} = G(P,Q).X(P,Q) + (1 - G(P,Q)).Q \tag{6c}$$

3.5. Deep Multimodal Fusion

Finally, the synthesized features from both self and gated cross interactions are further processed through a Bi-GRU to learn deep, integrative feature vectors for each modality. This stage consolidates the insights drawn from individual and cross-modal analyses, ensuring that the final feature representation is robust and comprehensive:

$$Deep_T = Bi\text{-}GRU(S_T, F_{VT}, F_{AT}) \tag{7}$$

This enriched multimodal feature vector is then inputted into a predictive layer, comprising a fully connected and softmax layer, to perform the final classification task, capturing the nuanced sentiments expressed across all modalities.

4. Experiments

4.1. Dataset

The ACMG model was evaluated using two widely recognized multimodal sentiment analysis datasets from the CMU Multimodal SDK: 1) CMU-MOSI: Multimodal Opinion Sentiment Intensity Dataset and 2) CMU-MOSEI: Multimodal Opinion, Sentiment, and Emotion Intensity Dataset. Both datasets are designed for binary sentiment classification, with sentiment values ≥ 0 indicating positive sentiments and values < 0 indicating negative sentiments. CMU-MOSI consists of 1284 training, 229 validation, and 686 test utterances, while CMU-MOSEI comprises 16216 training, 1835 validation, and 4625 test utterances.

4.2. Implementation Details

For our experiments with the ACMG model, we meticulously selected feature sets as recommended by Ghosal et al [34]. Specifically, for the CMU-MOSEI dataset, we utilized GloVe embeddings [74] to capture nuanced word-level features. Visual features were extracted using the Facets toolkit [1], renowned for its detailed analysis of machine learning model behavior. For acoustic features, we leveraged the robust capabilities of COVAREP [75], a cooperative voice analysis repository for speech technologies.

For the CMU-MOSI dataset, our approach incorporated a convolutional neural network (CNN) to derive utterance-level features. This was complemented by 3D CNN features for visual data and openSMILE [76] features for acoustic data, ensuring a comprehensive multi-modal data representation.

---

[1] https://pair-code.github.io/facets/

 doi:10.20944/preprints202408.0265.v1

Model training was carried out using Bi-directional Gated Recurrent Units (Bi-GRUs), with a hidden layer size of 100 for CMU-MOSI and 200 for CMU-MOSEI, adapting the network's complexity to the dataset's size. Regularization was implemented via a dropout rate of 0.4 to prevent overfitting, and ReLU activation functions [77] were employed to introduce non-linearity into the dense layers. Optimization was achieved using the Adam optimizer [78], with a learning rate of 0.0005. Batch sizes were set to 16 for CMU-MOSI and 32 for CMU-MOSEI, with the training process spanning 75 epochs to adequately converge on optimal solutions.

4.3. Results and Analysis

4.3.1. Baselines and Ablation Study

To rigorously evaluate the effectiveness of the ACMG model, we established several experimental conditions (Table 1). Initially, we defined a unimodal baseline (B1) and a bimodal baseline without gating (B3) to serve as fundamental comparisons. These baselines were crucial in understanding the incremental benefits introduced by self-attention (B2) and the gating mechanism (B4). Furthermore, the integration of these components into a deep multimodal fusion configuration (B6) was assessed.

Table 1. Performance comparison of ACMG model components showing accuracy improvements at each step.

Sl. No.	Model	CMU-MOSI	CMU-MOSEI
B1	Unimodal Baseline	80.57	78.58
B2	B1 + Self Attention	81.11	79.12
B3	Bimodal Baseline w/o Gating	81.91	80.00
B4	Bimodal Baseline w/ ACMG Gating	82.91	80.59
B5	B2 + B4 w/o Deep Multimodal Fusion	83.37	80.88
B6	**ACMG: Full Model w/ Multimodal Fusion**	**83.91**	**81.14**

The empirical results confirmed that incorporating self-attention improved the overall model performance by 0.54% on both the MOSI and MOSEI datasets. The introduction of the ACMG gating mechanism provided a further 1% improvement in accuracy on the MOSI dataset, demonstrating its effectiveness in managing modality-specific noise and enhancing feature integration. The deep multimodal fusion technique added an additional increase of 0.54% and 0.26% in accuracy on the MOSI and MOSEI datasets respectively, underscoring the synergy achieved through combined modality processing.

The performance improvements across these baselines substantiate our hypothesis that targeted, attention-driven interactions between modalities significantly enhance the model's robustness and sensitivity to context. The ACMG model's adept handling of long-term dependencies and its capability to integrate complex multimodal data into coherent feature representations underscore its advanced analytical prowess.

4.3.2. Benchmarking

To position the ACMG model within the current research landscape, we compared its performance against a variety of established benchmarks in multimodal sentiment analysis. These include: - Tensor Fusion Network [21], which amalgamates unimodal embeddings through a 3-fold Cartesian product; - Context-dependent Sentiment Analysis [20], focusing on dynamic multimodal feature extraction; - Memory Fusion Network (MFN) [32], which utilizes a sequential multi-view learning framework incorporating attention and memory; - Graph-MFN [30], which enhances the MFN architecture with a dynamic fusion graph for learning modality interactions; - Gated Multimodal Embedding with Temporal Attention [27], optimizing word-level modality fusion; - Hierarchical Fusion Approach [33], which executes sentiment analysis through a tiered fusion strategy at the word, sentence, and abstract

 doi:10.20944/preprints202408.0265.v1

levels; - Deep Canonical Correlation Analysis (DCCA) based Multi-modal Embeddings [22] for deep learning of joint modality representations.

In Table 2, the ACMG model demonstrates superior performance, surpassing state-of-the-art results by 1.6% on the CMU-MOSI corpus and 1.34% on the CMU-MOSEI corpus. This benchmarking not only highlights the model's efficacy but also its adaptability and precision in handling diverse and complex multimodal datasets.

Table 3 provides a qualitative insight into specific instances from the dataset, illustrating how the ACMG model selectively enhances modality-specific contributions and manages cross-modal interactions effectively, thereby yielding a high degree of accuracy in sentiment classification.

Table 2. Benchmark comparison of multimodal sentiment analysis performance across CMU-MOSI and CMU-MOSEI datasets. Asterisks indicate results on subsets excluding neutral sentiments.

CMU-MOSI			CMU-MOSEI		
Approach	**Accuracy**	**F1-Score**	**Approach**	**Accuracy**	**F1-Score**
Poria et al [20]	77.1	79.1	Zadeh et al [32]	76.0	76.0
Morency et al [19]	76.5	73.4	Poria et al [20]	76.9	77.0
Zadeh et al [33]	76.9	76.9	Morency et al [19]	77.64	-
Ghosal et al [34]	82.31	80.69	Ghosal et al [34]	79.80	-
Sun et al [22]	80.6	80.57	Sun et al [22]	(83.62)	(83.75)
ACMG Model	**83.91**	**81.17**	**ACMG Model**	**81.14 / (85.27)**	**78.53 / (84.08)**

Table 3. Detailed qualitative analysis of ACMG model's performance, showcasing the interplay of text, audio, and video modalities. S_{M_u} represents the self-attention score for each utterance u in the corresponding modality M. Cross-interaction scores are computed as the average values of the gating function $G(P, Q)$ for each pair of modalities P, Q.

Utterance	Gold Label	Predicted Label	Remark
Absolutely loved this experience	Pos.	Pos.	S_{T_u} = 0.95, indicating strong textual self-sufficiency Cross-interaction score: 0.45, showing moderate reliance on other modalities
Reflecting on how these performances were surprisingly overlooked at the Oscar awards this year	Pos.	Pos.	Despite the positive final sentiment, text analysis initially suggests neutrality. *T-A* and *T-V* interactions (0.18 and 0.07) indicate minor contributions. Visual (S_{V_u} = 0.78) and audio (S_{A_u} = 0.72) modalities primarily drive the positive outcome.
A few jokes were somewhat funny	Neg.	Neg.	Moderate correlation among all modalities, with cross-interaction scores of *T-A* = 0.82, *A-V* = 0.71, and *T-V* = 0.86, indicating a consistent interpretation across modalities.
I was completely blown away by the performance	Pos.	Pos.	Strong contributions from both audio (S_{A_u} = 0.65) and video (S_{V_u} = 0.53), with a high cross-interaction score of 0.79, emphasizing the impact of multimodal inputs.
The storyline was captivating, but some scenes dragged on	Mixed	Pos.	Text analysis shows mixed feelings (S_{T_u} = 0.60), while audio and visual modalities lean towards a positive sentiment (S_{A_u} = 0.85, S_{V_u} = 0.88). Cross-modal interaction scores, especially *T-V* (0.90) and *T-A* (0.87), highlight a compensatory effect of the other modalities.
The background score was fitting, though not memorable	Neu.	Neu.	Uniform agreement across all modalities with low cross-interaction scores (*T-A* = 0.30, *A-V* = 0.35, *T-V* = 0.32), indicating an aligned but subdued response from all inputs.

5. Conclusion and Future Work

In this paper, we introduced the Advanced Cross-Modal Gating (ACMG) model, a sophisticated approach designed to enhance multimodal sentiment analysis. The model innovatively combines self-attention mechanisms with a novel gating mechanism to optimize the integration and analysis of multimodal data. The self-attention component is crucial for capturing long-term contextual relationships within the data, while the gating mechanism effectively manages the integration of cross-modal interactions. This gating function is particularly adept at emphasizing relevant cross-modal interactions when unimodal cues are insufficient for accurate sentiment determination and downscaling their influence when unimodal information alone is robust enough to predict sentiments.

Our comprehensive evaluations on two benchmark datasets, CMU-MOSI and CMU-MOSEI, demonstrate that the ACMG model significantly outperforms existing state-of-the-art methods. The

 doi:10.20944/preprints202408.0265.v1

improvements observed underline the efficacy of our approach in handling complex, multimodal datasets by leveraging both the depth and nuances of multiple data types.

Looking forward, we aim to extend the capabilities of the ACMG model to more challenging real-world applications. One such domain is the analysis of customer interactions in call centers, where both text and audio modalities often suffer from significant noise due to poor recording quality and suboptimal speech recognition technologies. These conditions present unique challenges that our model, with its robust handling of noisy data and sophisticated attentional and gating mechanisms, is well-suited to address.

Moreover, we plan to explore the integration of additional modalities such as physiological signals and contextual metadata, which could provide deeper insights into the sentiments expressed during interactions. This expansion will likely involve the development of new gating mechanisms tailored to the specific characteristics and challenges of these data types.

Further research will also focus on improving the adaptability and efficiency of the ACMG model. This includes optimizing the model's architecture to reduce computational demands and enhance real-time processing capabilities, which are essential for applications such as live customer service interactions and on-the-fly content moderation.

In conclusion, the ACMG model represents a significant advancement in multimodal sentiment analysis, offering robust, adaptable, and accurate sentiment predictions. Its development not only addresses current challenges within the field but also sets the stage for future innovations that will extend its utility to a broader range of applications and data environments.

References

1. Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*, 2019.
2. Bing Liu. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.
3. Aniruddha Tammewar, Alessandra Cervone, and Giuseppe Riccardi. Emotion carrier recognition from personal narratives. *Accepted for publication at INTERSPEECH*, 2021. URL https://arxiv.org/abs/2008.07481.
4. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
5. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
6. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
7. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
8. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
9. F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1970–1973 vol.3, 1996.
10. Kang Liu, Heng Li Xu, Yang Liu, and Jun Zhao. Opinion target extraction using partially-supervised word alignment model. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.
11. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
12. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
13. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph

 doi:10.20944/preprints202408.0265.v1

Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, 2010. ISBN 2-9517408-6-7.

14. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
15. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
16. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
17. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
18. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
19. M. Wöllmer, F. Weninger, T. Knaup, B.W. Schuller, C. Sun, K. Sagae, and L.P. Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013. https://doi.org/10.1109/MIS.2013.34. URL https://doi.org/10.1109/MIS.2013.34.
20. Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 873–883, 2017.
21. A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.P. Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1103–1114, 2017. URL https://aclanthology.info/papers/D17-1115/d17-1115.
22. Z. Sun, P.K. Sarma, W. Sethares, and E.P. Bucy. Multi-modal sentiment analysis using deep canonical correlation analysis. *Proc. Interspeech 2019*, pages 1323–1327, 2019.
23. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
24. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
25. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
26. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
27. M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, and L.P. Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI*, pages 163–171, 2017. https://doi.org/10.1145/3136755.3136801. URL https://doi.org/10.1145/3136755.3136801.
28. A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, and L.P. Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
29. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
30. A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, and L.P. Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics, ACL, pages 2236–2246, 2018b. URL https://aclanthology.info/papers/P18-1208/p18-1208.

31. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
32. A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.P. Morency. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
33. E. Georgiou, C. Papaioannou, and A. Potamianos. Deep hierarchical fusion with application in sentiment analysis. *Proc. Interspeech 2019*, pages 1646–1650, 2019.
34. 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*, pages 3454–3466, 2018. URL https://aclanthology.info/papers/D18-1382/d18-1382.
35. S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *IEEE 16th International Conference on Data Mining, ICDM*, pages 439–448, 2016. https://doi.org/10.1109/ICDM.2016.0055. URL https://doi.org/10.1109/ICDM.2016.0055.
36. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
37. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
38. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
39. Timothy J Hazen, Shehzaad Dhuliawala, and Daniel Boies. Towards domain adaptation from limited data for question answering using deep neural networks. *arXiv preprint arXiv:1911.02655*, 2019.
40. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. 2024.
41. Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5801. URL https://aclanthology.org/D19-5801.
42. John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR, 2020.
43. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
44. Huimin Zeng, Zhenrui Yue, Yang Zhang, Ziyi Kou, Lanyu Shang, and Dong Wang. On attacking out-domain uncertainty estimation in deep neural networks. In *IJCAI*, 2022.
45. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12794–12802, 2021.
46. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13449–13467, 2023.
47. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
48. Tal Ben-Nun, Alice Shoshana Jakobovits, and Torsten Hoefler. Neural code comprehension: A learnable representation of code semantics. In *Advances in Neural Information Processing Systems*, volume 31, pages 3585–3597. Curran Associates, Inc, 2018.
49. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

 doi:10.20944/preprints202408.0265.v1

50. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 549–559, 2021.
51. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.
52. Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *The 30th International Conference on Machine Learning (ICML 2013)*, volume 28 of *JMLR Workshop and Conference*, pages 1139–1147, 2013.
53. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023.
54. Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023.
55. Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.
56. Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019.
57. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
58. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
59. Yang Zhang, Ruohan Zong, Jun Han, Hao Zheng, Qiuwen Lou, Daniel Zhang, and Dong Wang. Transland: An adversarial transfer learning approach for migratable urban land usage classification using remote sensing. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1567–1576. IEEE, 2019.
60. Yang Zhang, Ruohan Zong, and Dong Wang. A hybrid transfer learning approach to migratable disaster assessment in social media sensing. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 131–138. IEEE, 2020.
61. Yang Zhang, Daniel Zhang, and Dong Wang. On migratable traffic risk estimation in urban sensing: A social sensing based deep transfer network approach. *Ad Hoc Networks*, 111:102320, 2021.
62. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
63. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
64. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
65. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.
66. Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*, 2018.
67. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, 2023.
68. Feifan Fan, Yansong Feng, and Dongyan Zhao. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, 2018.

 doi:10.20944/preprints202408.0265.v1

69. Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. 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)*, pages 3570–3575, 2019.
70. K. Cho, B. Merrienboer, Ç Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1724–1734, 2014. URL http://aclweb.org/anthology/D/D14/D14-1179.pdf.
71. T. Luong, H. Pham, and C.D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015. URL http://aclweb.org/anthology/D/D15/D15-1166.pdf.
72. W. Wang, C. Wu, and M. Yan. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 1705–1714, 2018b. URL https://aclanthology.info/papers/P18-1158/p18-1158.
73. Y. Gong and S.R. Bowman. Ruminating reader: Reasoning with gated multi-hop attention. In *Proceedings of the Workshop on Machine Reading for Question Answering@ACL*, pages 1–11, 2018. URL https://aclanthology.info/papers/W18-2601/w18-2601.
74. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
75. G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP - A collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 960–964, 2014. https://doi.org/10.1109/ICASSP.2014.6853739. URL https://doi.org/10.1109/ICASSP.2014.6853739.
76. F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.
77. V. Nair and G.E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010. URL https://icml.cc/Conferences/2010/papers/432.pdf.
78. D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR*, 2015. URL http://arxiv.org/abs/1412.6980.