Article

Not peer-reviewed version

Enhanced Multimodal Integration Using TriFusion Networks for Comprehensive Emotion Analysis

Ethan Wilson , Rodolfo Patel , Ava Taylor * , Liam Jones

Posted Date: 8 August 2024

doi: 10.20944/preprints202408.0576.v1

Keywords: Emotion Recognition; Multimodal Fusion; Audio-Video Analysis

Article

Enhanced Multimodal Integration Using TriFusion Networks for Comprehensive Emotion Analysis

Ethan Wilson, Rodolfo Patel, Ava Taylor * and Liam Jones

Briar Cliff University
* Correspondence: taylor@briarcliff.edu

Abstract: In this work, we introduce the TriFusion Network, an innovative deep learning framework designed for the simultaneous analysis of auditory, visual, and textual data to accurately assess emotional states. The architecture of the TriFusion Network is uniquely structured, featuring both independent processing pathways for each modality and integrated layers that harness the combined strengths of these modalities to enhance emotion recognition capabilities. Our approach addresses the complexities inherent in multimodal data integration, with a focus on optimizing the interplay between modality-specific features and their joint representation. Extensive experimental evaluations on the challenging AVEC Sentiment Analysis in the Wild dataset highlight the TriFusion Network's robust performance. It significantly outperforms traditional models that rely on simple feature-level concatenation or complex score-level fusion techniques. Notably, the TriFusion Network achieves Concordance Correlation Coefficients (CCC) of 0.606, 0.534, and 0.170 for the arousal, valence, and liking dimensions respectively, demonstrating substantial improvements over existing methods. These results not only confirm the effectiveness of the TriFusion Network in capturing and interpreting complex emotional cues but also underscore its potential as a versatile tool in real-world applications where accurate emotion recognition is critical.

Keywords: emotion recognition; multimodal fusion; audio-video analysis

1. Introduction

The advent of sophisticated sensors capable of capturing high-fidelity audio and video data [50] has set the stage for breakthroughs in various fields, particularly in passive, non-invasive monitoring systems that could significantly enhance continuous healthcare management for chronic and mental health conditions including diabetes, hypertension, and depression [27]. The integration of these sensors into everyday environments like homes and offices is anticipated to revolutionize how spaces interact with occupants, adapting to and moderating their emotional and psychological states seamlessly and invisibly.

Emotion recognition has emerged as a pivotal area of research within affective computing, driven by the necessity to understand and interpret human emotions in a variety of applications ranging from interactive gaming to psychological analysis [31]. Recent advancements in this field have primarily leveraged deep learning techniques to enhance accuracy and efficiency in detecting emotions from complex datasets. Studies have increasingly focused on multimodal emotion recognition, integrating signals from various sources such as facial expressions, voice intonations, and physiological responses to achieve a holistic understanding of emotional states. Researchers like Kossaifi et al. have demonstrated the effectiveness of neural networks in disentangling these intricate modalities to predict emotions with greater precision [23,40]. Despite progress, challenges remain in handling the subtleties of context-dependent emotional expressions and the inherent subjectivity in emotional data interpretation, which continue to drive innovative solutions in this dynamic field.

The integration of audio and video data analysis is a critical aspect of multimodal emotion recognition, providing a richer context for understanding the nuances of human behavior. Audio-video analysis benefits from the confluence of visual cues, such as facial movements and body language, and auditory signals, like tone and pitch of voice, to form a comprehensive view of an individual's emotional state. The synchronization of these modalities presents unique challenges, particularly in aligning temporal dynamics and extracting meaningful features that are indicative of emotions.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2024 doi:10.20944/preprints202408.0576.v1

Pioneering work by Trigeorgis et al. on the fusion of audio and video streams through deep learning models exemplifies the advancements in this area, revealing the potential for significantly improved recognition rates over using single modalities [21,48]. These approaches underscore the necessity of developing robust algorithms that can efficiently process and analyze the complex interplay of auditory and visual data to enhance the accuracy and applicability of emotion recognition systems.

The potential of these technologies extends beyond mere convenience, aiming to provide critical support in managing conditions such as autism spectrum disorders, fatigue, and drug addiction through constant monitoring and immediate feedback. The capability to accurately identify and respond to affective states through multimodal emotional analysis is essential in realizing this future [11,21,22]. However, the journey towards effective real-world application is fraught with challenges, including the accurate capture and interpretation of complex spatio-temporal data across diverse populations and environmental conditions [45,46]. Additionally, the creation of expansive, well-annotated multimodal datasets necessary for training robust models remains a costly and labor-intensive endeavor.

To address these challenges, this paper proposes a novel approach to dynamic emotional state analysis using the TriFusion Network, which leverages deep learning to perform intermediate-level fusion of data from auditory, visual, and textual sources [1,3,10]. This method surpasses traditional early and late fusion techniques by optimizing feature extraction, classification, and fusion processes in a cohesive, end-to-end manner. The effectiveness of the TriFusion Network is rigorously validated against contemporary methodologies using the SEWA database, a benchmark in the field of affective computing.

Following this introduction, the paper is structured as follows: Section II provides a detailed review of existing emotional recognition methodologies across different modalities. Section III elaborates on the architecture and theoretical underpinnings of the proposed TriFusion Network. Section IV outlines the experimental setup employed to assess the network's performance, and Section V discusses the outcomes of these experiments across individual and combined modalities. The paper concludes with a summary of findings and a discussion on future research directions in Section VI.

2. Related Work

Over the last several years, the study of facial expression recognition (FER) has been propelled into the forefront of computational emotion analysis. Numerous methodologies have been developed to identify the seven universally recognized emotions—joy, surprise, anger, fear, disgust, sadness, and neutral—from static facial imagery [6,14,18,23,26,27]. These approaches are generally categorized into two primary techniques: appearance-based and geometric-based methodologies. The recent advent of dynamic facial expression recognition offers a compelling enhancement over static methods, analyzing emotions through a sequence of images or video frames which capture the temporal progression of facial expressions [19]. This dynamic approach facilitates the extraction of both spatial features and their temporal evolution, utilizing shape-based, appearance-based, and motion-based techniques for more nuanced emotion detection.

Shape-based methods, such as the Constrained Local Model (CLM), delineate facial structures using defined anchor points whose movements are tracked to infer emotional states. Appearance-based methods, exemplified by LBP-TOP, analyze textural and intensity patterns across facial images to classify expressions. Motion-based techniques, including free-form deformation models, examine the spatial-temporal dynamics of facial expressions, often necessitating robust facial alignment algorithms for accurate performance. For instance, Guo *et al.* utilized an atlas construction combined with sparse representation to concurrently harness spatial and temporal data, achieving significantly enhanced recognition accuracy by integrating these dimensions [9].

Emotion recognition has increasingly become a crucial field within human-computer interaction, facilitating advancements that range from customer service bots to therapeutic aids. Significant research has focused on improving recognition algorithms through machine learning models that process complex datasets from facial, vocal, and biometric modalities [72]. The field has seen a particular emphasis on the application of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture the subtle dynamics of emotional expressions over time [22]. Recent studies have explored the incorporation of context-aware systems that adjust their processing based on the situational context, aiming to tackle the variability and ambiguity inherent in human emotions [37]. These systems are designed not only to detect basic emotions but also to understand complex affective states and their transitions, challenging the traditional paradigms of emotion recognition with richer, more adaptive models [67–72].

In the realm of audio-video analysis for emotion recognition, the synergy between auditory and visual cues has been extensively studied to develop more accurate and reliable systems. This interdisciplinary approach leverages the strengths of each modality to compensate for the limitations of the other, often utilizing advanced signal processing and deep learning techniques [24,45]. the advancement of machine learning techniques, particularly supervised learning, For instance, the integration of facial expression analysis with voice tone analysis allows systems to discern subtleties in emotional expressions that might be ambiguous when only one modality is considered [101]. Researchers have developed frameworks that dynamically align audio and video data streams, extracting temporally correlated features to improve the coherence and accuracy of the emotion detection process [102,104]. These methodologies have been pivotal in advancing real-time emotion recognition systems, enhancing their application in real-world scenarios such as interactive media, surveillance, and telecommunication.

The automatic detection of emotional states through auditory cues has also seen significant advances, particularly in the context of depression and emotion detection. These systems draw parallels in their use of acoustic features to infer psychological states. Research by France *et al.* demonstrated that variations in formant frequencies could reliably indicate depression and suicidal tendencies [8]. Cummings *et al.* and Moore *et al.* achieved considerable success using energy, spectral, and prosodic features to classify depression with accuracies around 70-75% [7,16]. With the increasing prevalence of machine learning, deep neural networks, Long-Short Term Memory networks (LSTMs), and Convolutional Neural Networks (CNNs) have become ubiquitous in enhancing the precision of emotion detection systems [2,3,11,28].

The integration of multimodal data sources has been identified as a particularly potent method for improving the accuracy and reliability of emotion recognition systems. This approach is often implemented at feature, score, or decision levels, with each modality potentially providing complementary information that enhances overall system performance [2,3,11]. Recent studies have explored hierarchical frameworks that adaptively merge input modalities, leveraging varying degrees of certainty from vocal and facial data to detect depression and other emotional states [5,12]. For example, Meng *et al.* introduced a layered system utilizing Motion History Histogram features, and Nasir *et al.* employed a multi-resolution model combining audio and video features to diagnose depression more effectively [15,17]. Williamson *et al.* proposed a system that harnesses speech, prosody, and facial action units to assess depression severity, illustrating the value of multimodal integration [25].

Despite these advancements, several challenges persist in deploying these technologies in real-world scenarios. Often, models are built on limited datasets that may not be fully representative of the population [95,105,106], leading to potential biases and inaccuracies in emotion recognition. The operational variability in how data is captured—using standard equipment in uncontrolled environments—introduces additional complexity. The dynamic nature of human expressions and the contextual factors of recording environments necessitate adaptive models capable of handling intra-class variations and domain shifts. This paper proposes the use of the TriFusion architecture, a sophisticated deep learning model designed to effectively integrate multimodal information for robust emotion recognition. This approach extends beyond conventional feature-level and score-level fusion, implementing a hybrid system that optimizes both features and classifiers for comprehensive multimodal integration.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2024 doi:10.20944/preprints202408.0576.v1

3. Methodology

This paper introduces the TriFusion architecture, an innovative deep neural network (DNN) framework designed to robustly interpret the complex interplay of behavioral and emotional signals from multimodal sources. Leveraging the nuanced variations in facial expressions, vocal intonations, and textual cues captured in the AVEC SEWA database, TriFusion excels in learning optimal feature representations along with advanced classification and fusion strategies to predict emotional states such as arousal, valence, and liking accurately.

3.1. Feature Modeling

The core strategy of TriFusion involves the simultaneous learning of discriminative feature representations for each modality, accompanied by their respective classification and fusion into a cohesive decision-making framework. Initially, subsets of features from each modality—audio, video, and text—are processed independently through dedicated hidden layers. These layers are tailored to extract the most relevant features for the specific emotional recognition task at hand. Subsequently, the features are amalgamated in the later stages of the network through interconnected fully connected layers that execute both classification and fusion tasks, enabling the system to integrate and interpret multimodal data effectively.

3.1.1. Audio Features

In the audio domain, TriFusion processes 23 distinct acoustic low-level descriptors (LLDs), including energy, spectral components, cepstral features, pitch, voice quality, and micro-prosodic elements. These are sampled every 10ms across short-term frames. For each 6-second segment, a comprehensive feature vector is constructed using a codebook of 1,000 audio words, culminating in a 1,000-dimensional feature vector represented by a histogram of these audio words.

3.1.2. Video Features

Video data is handled by extracting key facial metrics at a frame rate of 20ms. This includes the normalized orientation of the face in degrees and the coordinates of critical facial landmarks—10 points around the eyes and 49 additional facial points. Each type of facial feature is encoded using a unique codebook, generating a histogram that contributes to a composite 3,000-dimensional feature vector for each video segment.

3.1.3. Text Features

Textual information is derived from transcriptions of spoken content, formatted into a bag-of-words model. The model encompasses 521 unique words, focusing solely on unigrams. These text-based features are aggregated over 6-second segments to form a feature set containing 521 distinct elements, providing a textual perspective on the expressed emotions.

3.2. TriFusion Architecture

The TriFusion architecture independently processes each modality—audio, video, and text—through dual-layer fully connected networks designed to capture intra-modality correlations. Following this, a concatenation layer merges these independent outputs into a unified representation, which is then fed into a subsequent fully connected layer that integrates the essence of all modalities. The final output is computed using a single linear neuron that functions as a regression estimator for the overall network, adjusted by a scaling module to align prediction magnitudes with actual label scales. Various scaling techniques such as decimal scaling, min-max normalization, and standard deviation scaling have been evaluated to optimize performance.

The training of the TriFusion model utilizes the Mean Squared Error (MSE) as its loss function, defined as follows:

 doi:10.20944/preprints202408.0576.v1

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \tag{1}$$

where $\hat{y}$, y, and m denote the predicted values, actual observed values, and the number of samples, respectively. During the training phase, the 'checkpoint' feature of the Keras API [4] is employed to save and retrieve the most effective model configuration, ensuring each dimension (arousal, valence, liking) is optimally trained for reliable performance.

3.3. System Training

3.3.1. Initialization

The training process of the TriFusion architecture begins with a rigorous initialization phase designed to set optimal conditions for learning. The initialization targets the setup of neural network parameters, specifically the weights and biases, which are crucial for the model's performance. We employ the He initialization method for weight setup, which is particularly effective for layers using ReLU activation functions by keeping the variance of activations across layers consistent. Each weight matrix W in the network is initialized according to the formula:

$$W = \sqrt{\frac{2}{n}} \cdot \mathcal{N}(0,1)$$

where n is the number of inputs to a layer, and $\mathcal{N}(0,1)$ denotes a standard normal distribution. Bias terms are initially set to zero, ensuring a neutral starting point for the first forward pass. This initialization phase is critical as it prevents the gradient vanishing or exploding problems commonly encountered in deep networks, thereby facilitating a stable and efficient gradient descent during the training phase.

3.3.2. Optimization and Backpropagation

Once initialization is complete, the TriFusion model enters the core phase of training, where it learns to minimize a predefined loss function through iterative optimization. We utilize the Adam optimizer, a method well-suited for large-scale and high-dimensional optimization problems. Adam combines the advantages of two other extensions of stochastic gradient descent, namely AdaGrad and RMSProp, specifically designed to handle sparse gradients on noisy problems. The optimizer adjusts the learning rate for each parameter based on estimations of first and second moments of the gradients:

$$\theta_{t+1} = \theta_t - \frac{\eta \cdot m_t}{\sqrt{v_t} + \epsilon}$$

where θ represents the parameters, η is the step size, m_t and v_t are estimates of the first and second moments of the gradients, respectively, and ϵ is a small scalar added to improve numerical stability. The backpropagation algorithm is applied to compute the gradient of the loss function with respect to each parameter in the network, effectively allowing the optimizer to update all weights and biases in the direction that minimizes the loss.

3.3.3. Loss Function and Regularization

The loss function is pivotal in guiding the training of the neural network. For the TriFusion system, which performs regression tasks, the Mean Squared Error (MSE) is used as the primary loss function, as defined:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 August 2024 doi:10.20944/preprints202408.0576.v1

where N is the number of training samples, $\hat{y}_i$ is the predicted value, and y_i is the actual value for the i-th sample. This choice ensures that the model is penalized based on the square of the difference between predicted and actual values, emphasizing larger errors more significantly, which is suitable for the regression nature of our task. Additionally, to combat overfitting—a common problem in deep learning architectures—L2 regularization is incorporated into the loss function. This regularization term adds a penalty equivalent to the square of the magnitude of coefficients, encouraging the model to maintain smaller weight values:

$$Loss = MSE + \lambda \sum_{w \in W} w^2$$

where λ is the regularization parameter, and W represents the set of all weights in the network. This regularization method not only helps in reducing overfitting but also promotes a more generalized model that performs well on unseen data.

These paragraphs comprehensively cover the setup, optimization, and regularization processes involved in training the TriFusion model, providing the necessary technical details and mathematical formulations to ensure clarity and depth in understanding the system's training dynamics.

4. Experimental Settings

4.1. Dataset and Protocol

The RECOLA dataset serves as the primary data source for our experiments, consisting of a training set and a development set. In our study, the development set was further divided into two subsets to refine the evaluation of the TriFusion model. The first subset comprises five subjects selected randomly from the original fourteen, while the remaining nine subjects form a secondary test subset. This division allows for a nuanced assessment of the model's generalization capabilities across different subsets of data.

For standard Support Vector Regression (SVR) approaches, the conventional protocol was adhered to for constructing unimodal and early fusion models. In contrast, our late fusion approach utilized the initial development subset for optimizing individual unimodal systems, with the fusion function refined using the secondary subset. This stratification addresses the need for precise model tuning under varied conditions.

4.2. Deep Neural Network Configurations

The TriFusion model's configuration varies according to the emotional dimension being evaluated—arousal, valence, and liking. Each dimension's specific architecture is meticulously designed to accommodate the idiosyncrasies of the corresponding emotional data. Table 1 details the DNN architecture for each emotional dimension, illustrating the modality-specific layer setups for audio, video, and text, along with the fusion layer's composition.

Table 1. TriFusion DNN architecture specified for each emotional dimension.

Modality	Arousal		Valence		Liking	
	L1	L2	L1	L2	L1	L2
Audio	50	50	200	200	50	50
Video	100	100	200	200	100	100
Text	200	200	200	200	100	100
Fusion Layer	100		200		50	

 doi:10.20944/preprints202408.0576.v1

The evaluation metric employed is the Concordance Correlation Coefficient (CCC), defined as:

$$\rho_{y'y} = \frac{2 * s_{y'y}}{s_{y'}^2 + s_y^2 + (\bar{y'} - \bar{y})^2}$$

where y' and y are the data sets for which the correlation is calculated, $s_{y'}^2$ and s_y^2 are the variances of these sets, and $\bar{y'}$ and $\bar{y}$ are their means.

4.3. Preprocessing Techniques

Adjustments for temporal delays significantly enhance CCC scores. An optimal delay value (d) was experimentally determined for each emotional dimension. For arousal and valence, $d = 1.5$ seconds was optimal, while for liking, extending the delay to $d = 2.5$ seconds provided the best results.

4.4. Postprocessing Strategies

Postprocessing in the TriFusion model involves scaling the DNN output to enhance prediction accuracy. We employ three scaling methods:

Min-Max Scaling Normalizes the output within a predefined range, calculated as:

$$\vec{y}_{norm} = \frac{(\max_l - \min_l) \cdot (\vec{y} - \min_p)}{\max_p - \min_p} + \min_l$$

where $\max_l$ and $\min_l$ are the maximum and minimum label values, respectively, and $\max_p$ and $\min_p$ are the maximum and minimum predicted values.

Standard-Deviation Ratio Adjusts predictions based on the ratio of standard deviations between predictions and labels, enhancing consistency across data scales:

$$\vec{y}_{norm} = \frac{\sigma_p}{\sigma_l} \otimes \vec{y}$$

where σ_p and σ_l are the standard deviations of predictions and labels, respectively, and $\otimes$ denotes element-wise multiplication.

Decimal Scaling Modifies the scale of prediction values to ensure they fall within a normalized range:

$$\vec{y}_{norm} = \frac{\vec{y}}{10^{\min_p}}$$

where $\min_p$ is the smallest power of ten for which the maximum absolute value of $\vec{y}_{norm}$ is less than one.

The implementation of these postprocessing techniques is critical for aligning the model's output with actual emotional states, ensuring both accuracy and reliability in the system's predictions.

5. Experimental Results and Discussion

This section delineates the experimental setup, discusses the methodologies employed, and analyzes the results derived from these experiments.

5.1. SVR-Based Baseline Results

Initial trials were conducted using Support Vector Regression (SVR) to establish a robust baseline for comparison. The parameters, including complexity, epsilon (ranging from 0.0 to 0.0001), and delay (ranging from 0 to 3 seconds), were meticulously optimized on the development set to ensure optimal performance. The results of these trials, as presented in Table 2, provide a comprehensive view of the performance across different modalities. The baseline paper protocol [20] was adhered to, with modifications made only to the early fusion configuration.

 doi:10.20944/preprints202408.0576.v1

A significant observation from these trials is the challenging nature of predicting 'liking' using audio and video modalities, whereas text data proved more efficacious. Despite the effectiveness of text, its reliance on transcription, which is both costly and time-consuming, poses a practical challenge for real-life applications. However, advancements in speech recognition technologies may potentially mitigate this issue. An intriguing aspect for future exploration is the impact of speech recognition inaccuracies on the precision of emotion detection systems.

Table 2. Performance of SVR models across different modalities and fusion techniques, measured by Concordance Correlation Coefficient (CCC).

Modality	Arousal			Valence			Liking		
	No Scaling	Std Ratio	Min-Max	No Scaling	Std Ratio	Min-Max	No Scaling	Std Ratio	Min-Max
Audio	.361	.400	**.449**	.412	.418	**.420**	.037	.028	**.040**
Video	.455	**.464**	.337	**.379**	**.379**	.344	**.174**	.166	.133
Text	.366	**.409**	.373	.380	**.402**	.399	**.315**	.301	.327
Early Fusion	.525	**.572**	.393	.508	**.532**	.491	.154	**.157**	.099
Late Fusion	**.387**	.358	.259	**.319**	.314	.398	.220	.247	**.290**

Late fusion experiments, utilizing simple linear regression techniques from the sklearn package, indicated varied effectiveness across dimensions. Notably, late fusion excelled for the liking dimension but did not perform as well in others compared to early fusion.

5.2. DNN-Based TriFusion Results

The deployment of the TriFusion DNN architecture was aimed at enhancing the integration of modal inputs more effectively than traditional methods. The early fusion DNN setup was directly contrasted with the SVR models to assess performance variations, with results detailed in Table 3. While the SVR and DNN models performed comparably for arousal, the SVR model showed a 6% improvement over the DNN for valence. Conversely, the DNN model demonstrated significant resilience against less effective modalities, outperforming the SVR by 26% for liking.

Table 3. Comparative performance of the early fusion DNN and SVR models on the development set, assessed via CCC.

Modality	Arousal			Valence			Liking		
	No Scaling	Decimal	Std Ratio	No Scaling	Decimal	Std Ratio	No Scaling	Decimal	Std Ratio
Early Fusion	.542	.542	**.565**	.467	.492	**.500**	.145	**.198**	.185
Proposed Fusion	.580	**.606**	.606	.530	.522	**.534**	.150	.165	**.170**

Interestingly, the TriFusion model surpasses both DNN and SVR models in predicting arousal and valence, though it still trails behind in the liking prediction when compared to the late fusion approach. This highlights potential areas for model refinement, particularly in managing detrimental modality effects. Ongoing and future investigations will focus on addressing these discrepancies and enhancing model robustness, especially given the initial promising results on the development set.

6. Conclusion and Future Work

6.1. Summary of Contributions

This research introduces the TriFusion architecture, a cutting-edge deep neural network (DNN) designed to enhance emotion recognition by integrating audio, video, and text modalities. This approach innovatively processes each modality through dual fully connected layers before merging them into a unified representation, effectively capturing the nuances of emotional states. Our end-to-end training methodology enables the TriFusion model to surpass previous architectures in terms of Concordance Correlation Coefficient (CCC) performance.

One of the key advancements of the TriFusion model is its ability to handle multimodal data seamlessly, ensuring robust feature extraction and fusion. Preliminary results have demonstrated that our model achieves superior performance metrics, suggesting that the detailed representation learning and fusion strategy are highly effective. However, there remains potential for further enhancement, particularly through the normalization of input features and the application of temporal smoothing techniques to stabilize the regressed outputs.

6.2. Technical Improvements and Optimization

Future developments will focus on refining the input normalization process to accommodate the dynamic range and distribution discrepancies across modalities. This refinement is expected to facilitate more consistent learning and improve the model's ability to generalize across diverse datasets. Additionally, implementing temporal smoothing on the output predictions will aim to reduce volatility and enhance the temporal coherence of the emotion recognition process, thereby aligning the predictions more closely with the inherent temporal progression of emotional states.

6.3. Expanding Model Capabilities

Looking ahead, the next phase of our research will explore the integration of a recurrent neural network (RNN) layer into the TriFusion architecture. This addition aims to capitalize on the temporal patterns in emotional expressions, potentially boosting accuracy and providing deeper insights into the sequential dynamics of emotions. By leveraging RNNs, we anticipate a significant improvement in the model's ability to track and predict emotional changes over time, which is crucial for applications requiring continuous emotional monitoring.

Furthermore, we plan to enhance the video component of our model by incorporating features extracted via a convolutional neural network (CNN) that is explicitly trained for emotion-related tasks. This approach is expected to refine the visual feature extraction process, allowing for more precise and contextually relevant information to be captured, which could dramatically improve the model's performance in real-world scenarios.

6.4. Broader Implications and Future Evaluations

The implications of these advancements extend beyond academic interest, promising significant applications in areas such as interactive media, teletherapy, and human-computer interaction, where understanding and responding to human emotions accurately is crucial. As part of our ongoing work, we will conduct extensive evaluations to assess the practical effectiveness of the TriFusion architecture across various domains and under different operational conditions.

Additionally, to ensure the robustness and applicability of our findings, future studies will involve cross-validation with larger, more diverse datasets and potentially real-time testing environments. These evaluations will help identify any biases or limitations in the current model and guide the development of more adaptive and resilient emotion recognition systems.

6.5. Conclusion

In conclusion, the TriFusion architecture represents a significant step forward in multimodal emotion recognition. By continuously refining and expanding this model, we aim to set new benchmarks in the field and contribute to the development of technologies that can empathetically interact with users across a spectrum of applications. Further research and development will be critical in realizing the full potential of this innovative approach.

References

1. A. Ali, N. Dehak, P. Cardinal, S. Khuranam, S. H. Yella, P. Bell, and S. Renals. Automatic dialect detection in arabic broadcast speech. In *Proc. of the 13th Annual Conf. of the Intl Speech Communication Association (Interspeech).*, 2016.

2. P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, and P. Boucher. ETS System for AV+EC 2015 Challenge. In *Proc. of the 5th Intl Workshop on Audio/Visual Emotion Challenge*, pages 17–23, New York, New York, USA, 2015.
3. S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proc. of the 5th Intl Workshop on Audio/Visual Emotion Challenge*, pages 49–56, 2015.
4. F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.
5. J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre. Detecting depression from facial actions and vocal prosody. In *3rd Intl Conf. on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7, Sept 2009.
6. M. J. Cossetin, J. C. Nievola, and A. L. Koerich. Facial expression recognition using a pairwise feature selection and classification approach. In *International Joint Conference on Neural Networks (IJCNN'2016)*, pages 5149–5155. IEEE, 2016.
7. N. Cummins, J. Epps, and E. Ambikairajah. Spectro-temporal analysis of speech affected by depression and psychomotor retardation. In *2013 IEEE Intl Conf. on Acoustics, Speech and Signal Processing*, pages 7542–7546, May 2013.
8. D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. on Biomedical Engineering*, 47(7):829–837, July 2000.
9. Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Trans. on Image Processing*, 25(5):1977–1992, May 2016.
10. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov 2012.
11. Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proc. of the 5th Intl Workshop on Audio/Visual Emotion Challenge*, pages 41–48, 2015.
12. M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. In *Proc. of the 3rd Intl Conf. on Pattern Recognition Applications and Methods*, pages 671–678, 2014.
13. B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proc. of Intl Conf. on Multimodal Interaction*, pages 427–434, New York, NY, USA, 2015.
14. J. Kumari, R. Rajesh, and K. Pooja. Facial expression recognition: A survey. *Procedia Computer Science*, 58:486 – 491, 2015. 2nd Intl Symposium on Computer Vision and the Internet.
15. H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proc. of the 3rd ACM Intl Workshop on Audio/Visual Emotion Challenge*, pages 21–30, October 2013.
16. E. Moore, M. Clements, J. Peifer, and L. Weisser. Analysis of prosodic variation in speech for clinical depression. In *Proc. of the 25th Annual Intl Conf. of the IEEE Engineering in Medicine and Biology Society*, volume 3, pages 2925–2928, Sept 2003.
17. M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proc. of the 6th Intl Workshop on Audio/Visual Emotion Challenge*, pages 43–50, 2016.
18. L. E. S. Oliveira, M. Mansano, A. L. Koerich, and A. S. Britto Jr. 2d principal component analysis for face and facial-expression recognition. *Computing in Science & Engineering*, 13(3):9–13, 2011.
19. M. Pantic and I. Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449, April 2006.
20. F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proc. of the 7th Intl Workshop on Audio/Visual Emotion Challenge*, Mountain View, USA, October 2017.
21. J. D. Silva Ortega, P. Cardinal, and A. L. Koerich. Emotion recognition using fusion of audio and video features. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1–6, 2019.

22. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Trans. on Medical Imaging*, 35(5):1299–1312, May 2016.
23. D. L. Tannugi, A. S. Britto Jr., and A. L. Koerich. Memory integrity of cnns for cross-dataset facial expression recognition. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1–6, 2019.
24. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *IEEE Intl Conf. on Acoustics, Speech and Signal Processing*, pages 5200–5204, March 2016.
25. J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proc. of the 4th Intl Workshop on Audio/Visual Emotion Challenge*, pages 65–72, 2014.
26. T. H. H. Zavaschi, A. S. Britto Jr., L. E. S. Oliveira, and A. L. Koerich. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, 40(2):646–655, 2013.
27. T. H. H. Zavaschi, A. L. Koerich, and L. E. S. Oliveira. Facial expression recognition using ensemble of classifiers. In *2011 ieee international conference on acoustics, speech and signal processing (ICASSP)*, pages 1489–1492. IEEE, 2011.
28. B. Zhang, C. Quan, and F. Ren. Study on cnn in the recognition of emotion in audio and images. In *IEEE/ACIS 15th Intl Conf. on Computer and Information Science*, pages 1–5, June 2016.
29. Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*, 2019.
30. Bing Liu. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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 Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, 2010. ISBN 2-9517408-6-7.
42. 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.

43. 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.
44. 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.
45. 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. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
46. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
47. 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. doi:10.1109/MIS.2013.34. URL https://doi.org/10.1109/MIS.2013.34.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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. doi:10.1145/3136755.3136801. URL https://doi.org/10.1145/3136755.3136801.
56. 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.
57. 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.
58. 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.
59. 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.

 doi:10.20944/preprints202408.0576.v1

60. 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.
61. E. Georgiou, C. Papaioannou, and A. Potamianos. Deep hierarchical fusion with application in sentiment analysis. *Proc. Interspeech 2019*, pages 1646–1650, 2019.

62. 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.
63. 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. doi:10.1109/ICDM.2016.0055. URL https://doi.org/10.1109/ICDM.2016.0055.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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. doi:10.18653/v1/D19-5801. URL https://aclanthology.org/D19-5801.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
78. 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.
79. 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.
80. 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.

81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*, 2018.
95. 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.
96. 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.
97. 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.
98. 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.

 doi:10.20944/preprints202408.0576.v1

99. 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.
100. 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.
101. 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.
102. 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.
103. 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. doi:10.1109/ICASSP.2014.6853739. URL https://doi.org/10.1109/ICASSP.2014.6853739.
104. 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.
105. 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.
106. 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.