Article

Not peer-reviewed version

Fusion-enhanced Multimodal Social Media Analysis

Ava Jackson , Rodolfo Patel , Ethan Taylor , Mia Anderson *

Posted Date: 12 August 2024

doi: 10.20944/preprints202408.0779.v1

Keywords: multimodal classification; social media analysis; emotion recognition

 doi:10.20944/preprints202408.0779.v1

Article

Fusion-Enhanced Multimodal Social Media Analysis

Ava Jackson, Rodolfo Patel, Ethan Taylor and Mia Anderson *

Briar Cliff University; mia.anderson@briarcliff.edu

Abstract: The analysis of social media content plays a crucial role in uncovering intricate user behaviors and trends across diverse digital platforms. Social media is inherently multimodal, incorporating texts, images, audios, and videos, each offering unique insights into user engagement and preferences. Traditional classification methods often focus narrowly on the most prominent modality, typically neglecting the synergistic potential of integrating multiple data types. To address this gap, we introduce the Unified Multimodal Classifier (UMC), a suite of streamlined models that adeptly harness and integrate these varied modalities. UMC leverages a novel architecture that combines a pooling layer with auxiliary learning tasks, facilitating the formation of a robust and shared feature space. This integration allows UMC to not only accommodate but also capitalize on the inherent diversity of social media data. The models are designed to be inherently flexible, adjusting to the availability of data modalities and maintaining high classification accuracy under varied conditions. In emotion classification scenarios, UMC has shown exceptional performance, significantly outperforming traditional methods by effectively synthesizing information across modalities. Its robustness is highlighted in its ability to deliver consistent results, even in the absence of one or more modalities. The simplicity and efficiency of UMC make it a potent tool for social media analytics, enabling it to achieve higher accuracies and maintain effectiveness even with limited data inputs. This adaptive capability ensures that UMC can be deployed in real-world applications where data incompleteness is common, thus broadening the scope of its applicability in various analytical contexts.

Keywords: multimodal classification; social media analysis; emotion recognition

1. Introduction

The proliferation of social media has led to a vast increase in user-generated content across various platforms. This expansion has transformed social media networks from specialized forums for sharing specific types of content, like texts or images, to versatile environments where users can share a diverse array of multimedia elements including texts, images, audio, and video clips. Current trends indicate that multimedia posts, particularly those incorporating images or videos, significantly enhance user engagement and interaction. Given the rich and expressive nature of multimodal content, it presents a unique opportunity to understand and predict user behaviors and preferences. By analyzing such multimodal data, we can extract valuable insights into user sentiments and emotions, which can be leveraged in numerous applications such as targeted advertising, personalized content delivery, and interactive marketing strategies. Despite the importance of accurate social media content classification, traditional methodologies often focus on a single modality, typically neglecting the rich context provided by combining multiple data types. For instance, while visual content may vividly portray emotions, textual content also carries nuanced emotional cues that are equally significant. By examining both text and images, we can achieve a more holistic understanding of the underlying sentiments in user posts.

The development of advanced sensors that capture high-quality audio and video data [79] has catalyzed significant advancements in various sectors, notably in the development of passive, non-invasive monitoring technologies. These technologies hold promise for improving ongoing management of chronic and mental health issues such as diabetes, hypertension, and depression [56].

 doi:10.20944/preprints202408.0779.v1

Expected to be incorporated into daily environments like homes and workplaces, these sensors could transform how these spaces interact with their inhabitants by subtly adjusting to and managing their emotional and psychological well-being.

Emotion recognition has become a crucial research focus within affective computing, propelled by the need to decode and interpret human emotions across diverse applications, from interactive gaming to psychological assessments [60]. The field has seen significant advancements through the application of deep learning methods, which enhance both the accuracy and the efficiency of emotion detection from complex data sets. Increasingly, research has turned toward multimodal emotion recognition, which combines insights from facial expressions, vocal tones, and physiological signals to form a comprehensive picture of an individual's emotional state. Innovations by researchers such as Kossaifi et al. have shown how neural networks can effectively untangle these complex inputs to predict emotions with increased accuracy [52,69]. However, the field continues to grapple with the nuances of context-specific emotional expressions and the subjective nature of interpreting emotional data, spurring ongoing innovation in this evolving area.

Analyzing both audio and video data is essential for multimodal emotion recognition, providing a deeper context for understanding human behavior's subtleties. This analysis benefits from merging visual indicators like facial expressions and body gestures with auditory elements such as voice tone and pitch, offering a detailed perspective on an individual's emotional state. The challenge of synchronizing these modalities involves matching their temporal dynamics and extracting relevant features indicative of emotional states. Groundbreaking efforts by Trigeorgis et al. in integrating audio and video data through advanced deep learning frameworks highlight significant progress in this domain, demonstrating potential for markedly improved recognition accuracy compared to single-modality approaches [50,77]. These developments emphasize the need for robust algorithms capable of effectively parsing and analyzing the intricate interactions between audio and visual data to boost the precision and usability of emotion recognition technologies.

These technologies offer more than just convenience; they aim to provide essential support in managing conditions like autism spectrum disorders, fatigue, and substance abuse through continuous monitoring and real-time feedback. The ability to accurately discern and react to emotional states via multimodal analysis is vital for these technologies to fulfill their potential [40,50,51]. Nonetheless, the path to effective deployment in real-world settings is laden with challenges, including the precise acquisition and analysis of complex spatio-temporal data across varied populations and conditions [74,75]. Moreover, developing extensive, well-annotated multimodal datasets for training effective models remains an expensive and time-consuming process.

To address these challenges, we propose the Unified Multimodal Classifier (UMC), which integrates neural network architectures to fuse multiple modalities effectively [30,32,39,71–73]. Unlike previous approaches that require the presence of all modalities and involve complex configurations tailored to specific tasks, UMC simplifies the integration process and enhances flexibility, facilitating application to a broader range of problems and accommodating the absence of certain modalities.

In this paper, we articulate our contributions as follows:

- We introduce a novel, generalized methodology, UMC, that amalgamates data from various modalities to improve the classification accuracy of social media content significantly.
- Our approach is designed to be robust, maintaining high classification performance despite the absence of one or more modalities, and is scalable to incorporate additional modalities or adapt to new application domains.
- We have developed and will release a comprehensive dataset that includes both textual and visual modalities, annotated with precise labels to facilitate extensive testing and future research in multimodal emotion analysis.

The remainder of the paper is structured as follows: Section 2 reviews existing literature in the fields of multimodal classification and emotion analysis. Section 3 details the traditional and

 doi:10.20944/preprints202408.0779.v1

our proposed fusion models. Section 4 elaborates on the specific methodologies employed in UMC. Experimental setups and results are discussed in Section 5, and Section 6 provides concluding remarks.

2. Related Work

The realm of multimodal classification in social media is broadly categorized into two distinct methodologies based on the integration of data from multiple sources. *Late fusion* processes modalities independently before combining the outcomes at the decision-making stage, operating under the assumption of modality independence, which often does not hold as different modalities usually depict correlated aspects of the same phenomena [22]. An innovative twist on late fusion utilizes the Kullback-Leibler divergence to align the results from different modalities, ensuring a more coherent decision process [27]. In contrast, *Early fusion* merges modalities at the data level, creating a unified feature set for subsequent classification [28]. This approach is favored in applications like sentiment analysis, where methods such as LSTM networks integrate visual and textual data [25], or hierarchical classifiers manage complex event categorization from combined features [28]. Beyond these, intermediate fusion strategies employ techniques like Latent Dirichlet Allocation (LDA) or Canonical Correlation Analysis (CCA) to discover underlying relationships between modalities in contexts such as image and text classification [11]. Although effective, these multimodal classification frameworks often require the presence of all modalities and can be quite complex. Our Unified Multimodal Classifier (UMC) introduces a simplified, yet robust, variant of early fusion that adapts to the absence of modalities and integrates divergent neural network architectures for different modalities, moving away from the traditional Siamese network configurations [12] where identical networks are used.

Emotion analysis techniques evolve from leveraging hand-crafted features derived from art and psychology to utilizing deep learning models that automatically extract discriminative features. Traditional methods involve low-level features such as shape [13], color, and texture [15], often combined into more complex representations [2,8,9,17]. These approaches, while intuitive, require extensive expert knowledge to design and may not capture all emotion-relevant aspects. To overcome these limitations, recent advancements have shifted towards deep learning, particularly using Convolutional Neural Networks (CNNs), to autonomously learn features from data [3]. Such techniques have demonstrated superior performance in emotion recognition solely from images. However, they neglect the rich emotional context provided by textual data. Our UMC model innovatively combines both visual and textual modalities using a hybrid CNN architecture, significantly enhancing the accuracy and depth of emotion analysis.

The importance of emotion recognition has grown within the domain of human-computer interaction, propelling forward innovations from customer service interfaces to therapeutic applications. There has been considerable focus on enhancing recognition algorithms using machine learning models that process intricate datasets from facial, vocal, and biometric signals [101]. The use of CNNs and Recurrent Neural Networks (RNNs) has been emphasized to capture the nuanced dynamics of emotional expressions over time [51]. Recent research has delved into context-aware systems that adapt their processing based on situational nuances, addressing the variability and ambiguity typical of human emotions [66]. These systems aim not just to identify basic emotions but to understand complex affective states and their fluctuations, challenging traditional emotion recognition models with more adaptive and enriched frameworks [96–101].

In the study of audio-video analysis for emotion recognition, the integration of auditory and visual cues has been extensively explored to develop more precise and reliable systems. This multidisciplinary approach utilizes the strengths of each modality to overcome the limitations of the others, frequently employing sophisticated signal processing and deep learning techniques [53,74]. For instance, combining facial expression analysis with voice tone analysis enables systems to discern emotional subtleties that might remain ambiguous when analyzed in isolation [130]. Researchers have crafted frameworks that dynamically align audio and video streams, extracting temporally

 doi:10.20944/preprints202408.0779.v1

correlated features to enhance the coherence and accuracy of emotion detection processes [131,133]. These methodologies are crucial in advancing real-time emotion recognition systems, expanding their practical applications in fields like interactive media, surveillance, and telecommunications.

The automatic detection of emotional states via auditory signals has also seen considerable advancements, particularly concerning depression and emotion recognition. These systems use acoustic features to infer psychological states, drawing parallels in their application. France *et al.* demonstrated that variations in formant frequencies could reliably indicate depression and suicidal tendencies [37]. Cummings *et al.* and Moore *et al.* have successfully used energy, spectral, and prosodic features to classify depression with accuracy rates around 70-75% [36,45]. The rise of machine learning has led to the widespread use of deep neural networks, Long-Short Term Memory networks (LSTMs), and Convolutional Neural Networks (CNNs) in refining the accuracy of emotion detection systems [31,32,40,57,93–95]

The integration of multimodal data sources has proven to be an effective method for improving the accuracy and reliability of emotion recognition systems. This approach typically involves combining features at the feature, score, or decision levels, with each modality providing complementary information that boosts the overall performance of the system [31,32,40]. Recent research has explored adaptive frameworks that intelligently merge input modalities, leveraging varying degrees of certainty from vocal and facial data to more accurately detect depression and other emotional states [34,41]. For instance, Meng *et al.* introduced a layered system that uses Motion History Histogram features, and Nasir *et al.* implemented a multi-resolution model that combines audio and video features for more effective depression diagnosis [44,46]. Williamson *et al.* developed a system that utilizes speech, prosody, and facial action units to assess the severity of depression, underscoring the value of multimodal integration [54].

Despite these advancements, challenges remain in deploying these technologies in real-world settings. Often, models are constructed using limited datasets that may not accurately represent the broader population, leading to potential biases and inaccuracies in emotion recognition [124,134,135]. The variability in data capture—often using standard equipment in uncontrolled environments—adds complexity to the process. The dynamic nature of human expressions and environmental factors necessitates adaptable models capable of handling variations within classes and shifting domains. This paper introduces the TriFusion architecture, an advanced deep learning model designed to effectively integrate multimodal information for robust emotion recognition. This approach goes beyond traditional feature-level and score-level fusion, implementing a hybrid system that optimizes both features and classifiers for comprehensive multimodal integration.

3. Multimodal Social Media Classification

This section outlines the various modalities encountered in social media, which are incorporated into our Unified Multimodal Classifier (UMC) models. Additionally, we review traditional and contemporary approaches for integrating these modalities within classification frameworks.

3.1. Multimodality in Social Media

Multimodality refers to the utilization of multiple communicative modes, including textual, aural, linguistic, spatial, and visual resources, to convey messages in social media [14]. Our research primarily focuses on the textual and visual modalities prevalent in social media platforms. The flexibility of our proposed UMC models allows for easy adaptation to include additional modalities if required.

A social media *post*, denoted as $x \in X$, may include an *image* (i), a *text* (s), or both, with an inherent semantic relationship assumed between text and image when both are present. Each image i is characterized by a feature vector $\gamma(i) \in \mathbb{R}^n$, which is derived using state-of-the-art convolutional neural networks (CNNs) trained on extensive image datasets. Textual content, ranging from short phrases to longer paragraphs, is represented by a vector $\psi(s) \in \mathbb{R}^m$, extracted via advanced text

 doi:10.20944/preprints202408.0779.v1

processing techniques such as word embeddings or transformer-based models, moving beyond traditional bag-of-words schemas.

3.2. Multimodal Classification using Feature Level Fusion

Our UMC approach aims to assign a post x to one of the classes in a set Y, based on the highest probability across potential classes,

$$\hat{y} = \arg\max_{y} P(Y = y|x). \tag{1}$$

For unimodal scenarios where x equals i or s, separate classifiers are trained for each modality. However, our focus is on the integration of both image and text modalities, utilizing traditional fusion techniques termed *late* fusion and *early* fusion, differentiated by the stage at which data integration occurs within the classification process.

3.2.1. Late Fusion

Late fusion involves the creation of two independent classifiers—one for images and one for texts. The final classification is determined by combining the output probabilities of these classifiers, where the predicted class maximizes the joint probability,

$$\hat{y} = \arg\max_{y} P(Y = y|i)P(Y = y|s). \tag{2}$$

This method assumes conditional independence between modalities, which simplifies the classification but may not fully capture the interdependencies between text and image content.

3.2.2. Early Fusion

Contrary to late fusion, early fusion amalgamates the modalities at the feature level before classification, thus leveraging the interplay between text and image features. This fusion is mathematically represented as,

$$\mathbf{x} = [\gamma(i); \psi(s)] \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^{(n+m)}$ represents the concatenated feature vector of post x. The combined vector $\mathbf{x}$ facilitates the application of sophisticated classification models such as support vector machines (SVM) or more complex neural networks. In our research, we employ deep learning techniques to construct the classifier on this enriched feature space, capitalizing on the strengths of both modalities to enhance classification performance.

4. Joint Fusion with Neural Network Models

While both late fusion and early fusion techniques integrate visual and textual data for classification, each method exhibits specific limitations. Late fusion, involving the creation of dual classifiers, often lacks efficiency, as the marginal gains from dual processing may not justify the complexity. Conversely, early fusion demands simultaneous availability of both image and text modalities, which might not always be feasible. In this section, we introduce two innovative approaches, termed *joint fusion* and *common space fusion*, which combine the simplicity of early fusion with the flexibility of late fusion under our Unified Multimodal Classifier (UMC) framework. These methods leverage neural networks for their adaptability and efficiency in learning directly from data, circumventing the need for extensive feature engineering.

 doi:10.20944/preprints202408.0779.v1

4.1. Mathematical Notations and Neural Layers

A neural network in our framework models the probability $P(Y = y|x)$ using a parametric function ϕ_θ (referenced in Equation 1), where θ encompasses all trainable parameters of the network. For an input x, the function ϕ_θ operates through multiple layers:

$$\phi_\theta(x) = \phi^L(\phi^{L-1}(\dots\phi^1(x)\dots)), \tag{4}$$

where L represents the total number of layers.

We denote matrices in bold uppercase letters (**X**, **Y**, **Z**) and vectors in bold lowercase (**a**, **b**, **c**). Element $\mathbf{A}_i$ refers to the i^{th} row of matrix **A**, and $[\mathbf{a}]_i$ denotes the i^{th} element of vector **a**. Vectors are assumed to be column vectors unless stated otherwise. Here, we describe two fundamental layers used in neural network-based classifiers: the *linear* layer and the *softmax* layer.

4.1.1. Linear Layer

This layer implements a linear transformation on its input **x**:

$$\boldsymbol{\phi}^l(\mathbf{x}) = \mathbf{W}^l\mathbf{x} + \mathbf{b}^l, \tag{5}$$

where $\mathbf{W}^l$ is the weight matrix and $\mathbf{b}^l$ represents the bias vector. The dimensions of these parameters adjust based on the input feature vector, whether it be from an image or text.

4.1.2. Softmax Layer

Following the computation of class scores in the penultimate layer, the softmax layer converts these scores into a probability distribution:

$$P(Y = y|x) \propto \phi_\theta(x, y) = \frac{\exp(\boldsymbol{\phi}^{L-1}(x, y))}{\sum_{k=1}^{|Y|}\exp(\boldsymbol{\phi}^{L-1}(x, y_k))} \tag{6}$$

4.2. Early Fusion as a Neural Network Model

Early fusion is depicted within our neural network model as a three-layer structure where the first layer, ϕ^1, combines the feature vectors from both modalities. For example: The fusion layer ϕ^1 performs a concatenation operation on the feature vectors $\gamma(i)$ and $\psi(s)$ from image and text respectively, resulting in the combined post vector $\mathbf{x} = [\gamma(i); \psi(s)]$. The subsequent layers, ϕ^2 and ϕ^3, are linear and softmax layers respectively, transforming the fused features into class probabilities. The learnable parameters in this setup are $\theta = \{\mathbf{W}^2, \mathbf{b}^2\}$, where these parameters are adjusted during training to minimize classification loss.

4.3. Unified Feature Representations

A critical component of our UMC is the robust representation of images and texts. Images are processed through a convolutional neural network to extract a feature vector $\gamma(i) \in \mathbb{R}^n$, utilizing pre-trained networks on large datasets like ImageNet for initial weights, which can be fine-tuned to our specific task. Texts are transformed into vectors $\psi(s) \in \mathbb{R}^m$ using embedding techniques followed by an aggregation layer that might employ operations like averaging or max pooling to condense word embeddings into a single text representation.

4.4. Joint and Common Space Fusion Models

Our joint fusion model is designed to handle cases where only one modality is available by modifying how image and text features are integrated. Rather than simple concatenation, this model employs pooling strategies to merge features into a unified representation, ensuring that either modality can independently contribute to the classification task.

 doi:10.20944/preprints202408.0779.v1

In common space fusion, the aim is to align the feature spaces of both modalities, enhancing the model's ability to learn from correlated features across different types of data. This alignment is facilitated by an auxiliary learning task that optimizes the feature representations to be similar for the same class and distinct across different classes, thus refining the model's generalization capabilities across diverse social media posts.

5. Experiments with Emotion Classification

This section presents a comprehensive evaluation of our newly proposed Unified Multimodal Classifier (UMC) on the task of emotion classification using both visual and textual data. To our knowledge, this novel approach is pioneering in its field.

5.1. Emotion as Discrete Categories

In the realm of emotion classification, our objective is to assign a given social media post x to an appropriate emotion class from a predefined set $Y = \{y_1, ..., y_k\}$, where each class y_i is mutually exclusive. We adopt Plutchik's renowned model of primary emotions as the foundation for our classification schema [21].

5.2. Datasets

There is a notable absence of large-scale datasets comprising both visual and textual data specifically curated for emotion classification. This scarcity prompted us to both enhance an existing dataset and construct a new one from the ground up.

Enriching an image-only dataset Originally compiled by You et al. [26], the *flickr* dataset contains images tagged according to eight emotional categories by multiple annotators from Amazon Mechanical Turk. We expanded this dataset by scraping titles and descriptions from Flickr, ensuring all textual data is in English and contains more than five words. The updated *flickr* dataset statistics are presented in Table 1.

Constructing an emotion dataset from scratch We also developed a dataset by aggregating content from various Reddit subreddits associated with specific emotions, such as *happy* (joy), *creepy* (fear), *rage* (anger), and *gore* (disgust). Each selected submission included both image and corresponding textual content, ensuring rich multimodal data. The collection process targeted posts with a high number of upvotes to ensure relevance and engagement.

Table 1. Detailed statistics of the enriched *flickr* and newly created Reddit datasets for emotion classification.

Dataset	Emotion	Number of Posts
Reddit	Joy	1119
	Fear	697
	Anger	613
	Disgust	810
Flickr	Amusement	1259
	Anger	407
	Awe	1561
	Contentment	2389
	Disgust	852
	Excitement	1451
	Fear	434
	Sadness	984

5.3. Experimental Setup

Our experimental framework employs the GloVe word vectors trained on Twitter data, aligning with the social media context of our datasets. We chose a word vector size of 200 for optimal

 doi:10.20944/preprints202408.0779.v1

performance, as validated against other vector sizes. Each model's performance was rigorously evaluated against a baseline of traditional and single-modality classifiers on both datasets, as shown in Table 2.

Table 2. Comparative performance of UMC against traditional single-modality and other multimodal fusion methods on the Reddit and Flickr datasets.

Model	Reddit			Flickr		
	Accuracy	F-Macro	F-Micro	Accuracy	F-Macro	F-Micro
Image-based	77.5%	0.77	0.78	56.9%	0.67	0.73
Text-based	78.4%	0.77	0.78	88.3%	0.85	0.88
Late fusion	83.3%	0.82	0.83	91.8%	0.87	0.91
Early fusion	84.1%	0.81	0.83	92.7%	0.89	0.92
UMC - Joint	86.3%	0.84	0.86	93.0%	0.90	0.93
UMC - Common Space	**86.9%**	**0.85**	**0.87**	**93.4%**	**0.91**	**0.934**

5.4. Results and Discussion

The results demonstrate that UMC consistently outperforms all baselines and traditional fusion methods, confirming the effectiveness of integrating visual and textual modalities for emotion classification. Specifically, the Common Space Fusion approach of UMC shows a slight advantage over Joint Fusion, suggesting that creating a shared feature space for different modalities enhances the classifier's ability to generalize across diverse inputs.

5.5. Qualitative Analysis

Further qualitative analysis reveals specific instances where UMC correctly identifies complex emotions from multimodal inputs, highlighting scenarios where either visual or textual cues alone might lead to misclassification. This analysis substantiates the robustness of UMC in handling real-world, noisy social media data.

6. Conclusions and Future Work

In this study, we introduced the Unified Multimodal Classifier (UMC), a set of models designed to efficiently integrate data from various modalities to perform social media analysis. The UMC framework is designed to be highly adaptable, capable of processing inputs even when some modalities are absent, and demonstrates its robustness across different scenarios. Our experiments on emotion classification reveal that UMC, while straightforward in its architecture, delivers impressive performance, surpassing traditional unimodal and multimodal approaches. We supported our research with two custom-constructed multimodal datasets, which were instrumental in demonstrating the efficacy of our models under diverse conditions.

6.1. Contributions

The primary contributions of this work are: Development of the UMC, which simplifies the integration of multimodal data for social media analysis. Validation of UMC's effectiveness in handling incomplete data modalities without compromising on performance. Creation of two novel datasets that encompass a wide range of emotions, providing a comprehensive platform for testing multimodal emotion classification.

6.2. Future Directions

Looking forward, we aim to expand the capabilities of UMC by exploring additional modalities such as structured and user-annotated data, which hold the potential to enrich the models' understanding and accuracy further [7,19]. Moreover, while this research focused primarily on

 doi:10.20944/preprints202408.0779.v1

emotion classification, future applications could include but are not limited to, sentiment analysis, behavioral prediction, and personalized content delivery.

6.3. Challenges and Opportunities

The integration of increasingly complex modalities presents both challenges and opportunities. Challenges include the scalability of the models to handle large-scale data and the ability to maintain high accuracy levels across diverse datasets. On the other hand, opportunities lie in leveraging cutting-edge AI techniques like deep learning and natural language processing to enhance the interpretability and efficiency of multimodal data analysis.

6.4. Extending to Real-World Applications

We also plan to collaborate with industry partners to test the real-world applicability of UMC in various sectors such as marketing, healthcare, and public safety. These applications could benefit from advanced predictive analytics, offering insights that are not only accurate but also actionable.

In summary, the UMC framework sets a new benchmark for multimodal data fusion in social media analytics. Its ability to handle incomplete data, combined with high accuracy and simplicity of use, makes it a promising tool for researchers and practitioners alike. The continued development and application of UMC will undoubtedly open new avenues in the field of artificial intelligence and data science.

References

1. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N.d., Blei, D.M., Jordan, M.I.: Matching words and pictures. JMLR 3(Feb) (2003)
2. Borth, D., Chen, T., Ji, R., Chang, S.F.: Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In: MM (2013)
3. Chen, M., Zhang, L., Allebach, J.P.: Learning deep features for image emotion classification. In: ICIP. IEEE (2015)
4. Ciresan, D.C., Meier, U., Masci, J., Maria Gambardella, L., Schmidhuber, J.: Flexible, high performance convolutional neural networks for image classification. In: IJCAI (2011)
5. Cui, B., Tung, A.K., Zhang, C., Zhao, Z.: Multiple feature fusion for social media applications. In: SIGMOD. ACM (2010)
6. Hoffer, E., Hubara, I., Ailon, N.: Deep unsupervised learning through spatial contrasting. arXiv preprint arXiv:1610.00243 (2016)
7. Hung, N.Q.V., Thang, D.C., Weidlich, M., Aberer, K.: Minimizing efforts in validating crowd answers. In: SIGMOD. pp. 999–1014 (2015)
8. Irie, G., Satou, T., Kojima, A., Yamasaki, T., Aizawa, K.: Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. TMM (2010)
9. Jia, J., Wu, S., Wang, X., Hu, P., Cai, L., Tang, J.: Can we understand van gogh's mood?: learning to infer affects from images in social networks. In: MM (2012)
10. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
11. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. TPAMI 29(6) (2007)
12. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop. vol. 2 (2015)
13. Lu, X., Suryanarayan, P., Adams Jr, R.B., Li, J., Newman, M.G., Wang, J.Z.: On shape and the computability of emotions. In: MM (2012)
14. Lutkewitte, C.: Multimodal composition: A critical sourcebook. Bedford/St. Martin's (2013)
15. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: MM (2010)

 doi:10.20944/preprints202408.0779.v1

16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
17. Nguyen, Q.V.H., Duong, C.T., Nguyen, T.T., Weidlich, M., Aberer, K., Yin, H., Zhou, X.: Argument discovery via crowdsourcing. VLDBJ (2017)
18. Nguyen, T.T., Duong, C.T., Weidlich, M., Yin, H., Nguyen, Q.V.H.: Retaining data from streams of social platforms with minimal regret. In: IJCAI (2017)
19. Nguyen, T.T., Nguyen, Q.V.H., Weidlich, M., Aberer, K.: Result selection and summarization for web table search. In: ICDE. pp. 231–242 (2015)
20. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP (2014)
21. Plutchik, R.: The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. AS (2001)
22. Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Audiovisual synchronization and fusion using canonical correlation analysis. IEEE Trans. Multimedia (2007)
23. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: CVPR (2014)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
25. You, Q., Cao, L., Jin, H., Luo, J.: Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In: MM (2016)
26. You, Q., Luo, J., Jin, H., Yang, J.: Building a large scale dataset for image emotion recognition: The fine print and the benchmark. arXiv preprint arXiv:1605.02677 (2016)
27. You, Q., Luo, J., Jin, H., Yang, J.: Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: WSDM (2016)
28. Zeppelzauer, M., Schopfhauser, D.: Multimodal classification of events in social media. Image and Vision Computing 53 (2016)
29. Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.S., Sun, X.: Exploring principles-of-art features for image emotion recognition. In: MM (2014)
30. 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.
31. 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.
32. 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.
33. F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.

 doi:10.20944/preprints202408.0779.v1

40. 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.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*, 2019.
59. Bing Liu. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.
60. 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.

 doi:10.20944/preprints202408.0779.v1

61. 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.
62. 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.
63. 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.
64. 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.
65. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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://aclanthology.org/N19-1423.
75. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
76. 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.
77. 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.
78. 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. https://aclanthology.info/papers/D17-1115/d17-1115.
79. 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.

 doi:10.20944/preprints202408.0779.v1

80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.
90. E. Georgiou, C. Papaioannou, and A. Potamianos. Deep hierarchical fusion with application in sentiment analysis. *Proc. Interspeech 2019*, pages 1646–1650, 2019.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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*

 doi:10.20944/preprints202408.0779.v1

for Question Answering, pages 1–13, Hong Kong, China, November 2019. Association for Computational Linguistics. https://aclanthology.org/D19-5801.

99. 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.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. 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.
106. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. 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.
113. 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.
114. 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.
115. 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.
116. 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.
117. 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.

 doi:10.20944/preprints202408.0779.v1

118. 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.
119. 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.
120. 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.
121. 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.
122. 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.
123. Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*, 2018.
124. 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.
125. 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.
126. 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.
127. 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.
128. 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.
129. 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.
130. 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.
131. 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.
132. 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.
133. F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia featurer extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.
134. 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.

 doi:10.20944/preprints202408.0779.v1

135. 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.