Academic Education Publishing House

A Survey of Speech and Text-Based Emotion Recognition

Zhe Liu, Ruiyu Liu
Maynooth international college, Fuzhou university, Fuzhou, Fujian, China

Abstract: Affective recognition serves as a core technology for anthropomorphic interaction in artificial intelligence. Traditional single-modal approaches (relying solely on speech or text) are limited by issues such as incomplete feature representation and environmental sensitivity. Multimodal sentiment analysis, which integrates speech prosody (e.g., pitch, speech rate) and text semantics (e.g., emotional vocabulary, syntactic structures), significantly enhances recognition accuracy and robustness. This paper systematically reviews the research progress in this field.

Keywords: Emotion Analysis; Multimodal Sentiment Analysis; Deep Learning; Multimodal Fusion; Emotion Representation Model; Cross-Domain Generalization

1. Introduction

Affective computing represents a core capability for artificial intelligence to achieve anthropomorphic interaction. Conventional single-modal approaches (e.g., relying solely on speech or text) suffer from issues such as incomplete feature representation and environmental interference sensitivity [1]. The complementary nature of speech and text provides a more accurate methodology for emotion recognition: speech signals contain paralinguistic features including prosody and speech rate, while textual content includes emotional lexicon and semantic orientation. The multimodal fusion of these two modalities can overcome the limitations of unimodal analysis. Additionally, speech and text share unified information acquisition sources. With the advancement of deep learning technologies, speech-text dual-modal emotion recognition has exhibited significant application value in fields such as voice assistants, intelligent customer service, and mental health assessment [2].

2. Emotion Representation Model

Emotion modeling serves as the core of dual-modal emotion analysis, aiming to transform complex human emotional expressions into computational forms. According to different modeling approaches, emotion models can be classified into four categories: Discrete Emotion Model: Suitable for simple and explicit emotion classification tasks and annotation cost-sensitive tasks; Dimensional Emotion Model: Appropriate for continuous emotion prediction and multimodal data fusion; Three-dimensional Emotion Model: Mainly used for complex emotional state modeling; Deep Learning Model: Applied to complex scenario modeling and multimodal fusion tasks.

2.1 Categorical Models

Based on Ekman's "Basic Emotion Theory", it is hypothesized that human emotions can be decomposed into a finite set of discrete categories (Anger, Disgust, Fear, Joy, Sadness, Surprise). By assigning explicit emotional labels, emotion analysis is transformed into a supervised classification task. This categorical approach offers advantages of labeling simplicity and interpretability, yet inherently overlooks the continuous and nuanced nature of human affect.

2.2 Dimensional Models

Dimensional models conceptualize emotions as continuous variation systems, describing and classifying affective states according to orthogonal dimensions. Prominent frameworks include two-dimensional (e.g., valence-arousal) and three-dimensional (e.g., valence-arousal-dominance) emotion models.

The two-dimensional emotional space model proposed by Russell uses valence (degree of pleasure/displeasure) and arousal (degree of activation/deactivation) as orthogonal axes to partition emotions into four quadrants. Quadrant I represents positive valence with high arousal, associated with joy; Quadrant III reflects low arousal and negative valence, linked to sadness; Quadrant II is characterized by high arousal and negative valence, typified by anger; Quadrant IV exhibits low arousal and positive valence,

corresponding to calmness.

2.3 Three-Dimensional Emotion Model

2.3.1 The PAD three-dimensional emotion theory proposed by Mehrabian and Russell in 1974 posits that emotions are characterized by three orthogonal dimensions: pleasure (valence), arousal (activation level), and dominance (perceived control). Pleasure reflects the hedonic quality of emotional states, indicating the degree of positivity or negativity; arousal represents neurophysiological activation levels and alertness; dominance measures an individual's perceived control over situations and others, reflecting subjective control over emotional states.

2.3.2 The three-dimensional emotion theory first proposed by Wundt [3] in 1896 posits that emotions require three orthogonal dimensions—pleasure-displeasure, tension-relaxation, and excitement-calmness—to achieve effective characterization. Each specific emotion occupies a distinct position between the two poles of these three dimensions, with each affective state representing a unique combination of these dimensions.

2.3.3 Schlosberg's 1950s three-dimensional emotion theory arranges specific emotions on an inverted elliptical cone using three orthogonal dimensions: pleasure-displeasure, attention-rejection, and activation level. The major axis of the elliptical section represents the hedonic dimension, the minor axis represents the attention dimension, and the vertical axis perpendicular to the elliptical plane represents the intensity of activation. Diverse emotional states arise from different combinations of these three-dimensional levels.

2.3.4 Plutchik's three-dimensional emotion model posits three dimensions: intensity, similarity, and polarity. He employs an inverted cone structure to illustrate the relationships among these dimensions. The cone's cross-section is divided into eight primary emotions, where adjacent emotions share similarity and diametrically opposed emotions represent polar opposites. The vertical axis of the cone reflects the gradient of emotional intensity from weak (base) to strong (apex).

Dimensional models offer greater flexibility in describing emotional dynamics, making them well-suited for scenarios requiring nuanced analysis. However, they are characterized by high labeling costs and the need to address data sparsity issues.

2.4 Deep Learning Model

Deep Learning Models automatically learn multimodal emotional features through neural network architectures (e.g., Recurrent Neural Networks, Transformers, Graph Neural Networks), making them suitable for large-scale, high-dimensional data. Their core strength lies in end-to-end feature extraction, eliminating the need for manual feature engineering and enabling the capture of complex semantic, prosodic, and nonlinear inter-modal relationships.

Table 1. Comparison of Four Emotion Models

Decision Dimension	Discrete Model	Dimensional Model	Three-Dimensional	Deep Learning Model Model
Task Type	Classification tasks (such as customer service emotion classification)	Continuous prediction (such as real-time monitoring)	Complex emotion analysis (such as psychological assessment)	Multimodal fusion (such as social media analysis)
Data Scale	Small sample size (such as clinical dialogue data)	Medium sample size (such as the SEMAINE dataset)	Small sample size (psychological experiment data)	Large scale (such as the CMU-MOSEI dataset)
Interpretability Requirement	High (such as medical diagnosis)	Medium (such as advertising effect analysis)	Low (theoretical research)	Low (black-box applications in the industry)
Deployment Cost	Low (lightweight model)	Medium (requiring continuous calculation)	High (requiring support from professional tools)	High (requiring GPU acceleration)

3. Analysis of Current Research Status on Speech and Text-Based Dual-Modal Emotion Recognition

Current approaches to speech-text dual-modal emotion recognition primarily focus on feature fusion, with early fusion and attention mechanisms serving as key methods[4]. These methods have achieved high accuracy rates (70%+) on datasets such as IEMOCAP[5]. However, challenges remain in modal alignment, data labeling, and model lightweighting. Future

research should explore self-supervised learning, cross-modal alignment algorithms, and ethical compliance frameworks to facilitate technological implementation in real-world scenarios like intelligent customer service and mental health monitoring.

3.1 Speech Feature Extraction

Speech emotion analysis relies on the extraction and modeling of acoustic features, primarily including:

3.1.1 Traditional Acoustic Features:

(1) Prosodic features include pitch (fundamental frequency), speech rate, energy (amplitude), and pauses, which reflect emotional intensity (e.g., faster speech rate and higher pitch during anger)[6].

(2) MFCC (Mel Frequency Cepstral Coefficients): Derived from the human auditory system's characteristics, this method converts the Mel-frequency spectrum into cepstral parameters, widely used in speech recognition and emotion analysis[6, 7].

(3) Mel spectrogram: Frequency features are extracted using a Mel filter bank to capture emotion-related timbral variations[6].

3.1.2 Deep Learning-Based Features" 或 "Deep-Learned Features:

(1) CNN: Trigeorgis [8] proposed AdieuNet, a CNN-based model that automatically learns speech features and achieved state-of-the-art (SOTA) performance on the IEMOCAP dataset[5].

(2) LSTM/GRU: Wollmer [9] incorporated LSTM (Long Short-Term Memory) with emotional labels to model temporal dynamics, such as periodic variations in anxious speech.

(3) Pretrained models: For example, HuBERT[10] employs self-supervised learning to extract high-level features from unlabeled speech, enhancing cross-domain generalization capabilities[2].

3.2 Text Feature Extraction

Text features serve as the foundation for sentiment analysis, primarily categorized into semantic, syntactic, and contextual features:

3.2.1 Word Embedding and Semantic Representation:

(1) Static Word Embedding: Models like GloVe [11] (Global Vectors for Word Representation) learn word vectors through global word co-occurrence statistics to capture semantic similarity.

(2) Dynamic Contextual Embedding: BERT (Bidirectional Encoder Representations from Transformers) generates context - related word vectors through bidirectional Transformer pre - training, significantly improving the performance of complex semantic tasks [1].

(3) Emotion Lexicon: VADER [12] (Valence Aware Dictionary and sEntiment Reasoner) quantifies the emotional polarity of text based on rules and a lexicon, making it suitable for rapid analysis of short texts [3].

3.2.2 Syntactic and structural features: These rely on syntactic parsing (e.g., Stanford CoreNLP [13]) to extract sentence components like subject-verb-object (SVO) and capture emotion-triggering structures such as negation and contrast. Dependency tree-based recursive neural networks integrate syntactic hierarchy into emotion analysis.

3.2.3 Contextual models: BiLSTM (Bidirectional Long Short-Term Memory) [14] captures long-range dependencies and integrates attention mechanisms [15] to focus on critical sentiment words.

3.2.4 Transformer: The self-attention mechanism proposed by Vaswani [16] exhibits superior performance in long-sequence modeling, such as in dialogue emotion analysis [3].

3.3 Commonly Used Datasets

3.3.1 IEMOCAP (Interactive Emotional Dyadic Motion Capture Database). Released in 2008 by the SAIL Lab at the University of Southern California (USC)[5], IEMOCAP is one of the earliest multimodal emotion dialogue datasets. The dataset contains 151 dyadic dialogues (302 videos) from 10 actors, spanning approximately 12 hours across 5 sessions. It is commonly used for multimodal emotion recognition and emotional modeling in dialogue systems [17]:

Text: Dialogue transcripts segmented into utterances; Audio: 48kHz sampling rate, including speech and non-verbal sounds (e.g., laughter); Vision: 53 facial motion capture markers (FACS - Facial Action Coding System) recording facial expressions, head movements, and gestures.

Discrete emotions: anger, happiness, sadness, neutral, excitement, frustration, etc., totaling nine categories.

Continuous attributes: Arousal, Valence, Dominance.

3.3.2 CMU-MOSI/MOSEI (Multimodal Opinion Sentiment and Emotion Intensity).

CMU-MOSI: Released in 2016, containing 2,199

YouTube lecture videos.
CMU-MOSEI: Extended in 2018 with 23,453 videos, making it one of the largest multimodal emotion datasets [1].
Suitable for large-scale multimodal model training and cross-domain emotion analysis [2].
Text: Transcribed text using GloVe word embeddings; Audio: 74-dimensional acoustic features extracted by COVAREP (e.g., Mel-frequency cepstral coefficients (MFCC), fundamental frequency); Vision: 35-dimensional facial action units (FAUs) extracted by Facet.
Emotion categories: six basic emotions according to Ekman (happiness, sadness, anger, etc.).
Emotional intensity: continuous ratings from -3 (strong negative) to +3 (strong positive).
3.3.3 SEMAINE (Sensitive, Expressive, and Affective Interactions in a Multimodal Empathic System). Released in 2012 by the University of Sheffield, UK, and collaborating institutions [18], SEMAINE is suitable for emotion analysis in human-computer dialogue, continuous emotion prediction (e.g., AVEC Challenge), and affective computing model development. The dataset contains 150 human-computer dialogue videos (approximately 6 hours total) partitioned into training, validation, and test sets [3]:
Text: Dialogue transcripts; Audio: 16kHz sampling rate, including speech and non-verbal sounds; Vision: Facial expressions and head movements.
Continuous dimensions: Arousal, Valence, Dominance, annotated every 0.2 seconds.

3.4 Multimodal Fusion Methods

Multimodal Fusion Methods are a critical research direction in the field of multimodal information processing, aiming to integrate information from speech and text modalities to enhance system performance and accuracy.
3.4.1 Early Fusion-Feature Concatenation: The feature concatenation method in early fusion was proposed by Poria et al. in 2016 [19]. This approach concatenates text word vectors with speech MFCC features and inputs them into unified models such as LSTM or CNN for processing [20]. By leveraging interactions between low-level features, this method enables cross-modal information to influence each other at an early stage, thereby improving model performance. For example, in speech emotion recognition, concatenating speech MFCC features with text description features and inputting them into a support vector machine (SVM) can enhance emotion classification accuracy. However, this method has limitations, primarily due to high dimensionality and computational complexity. High-dimensional feature vectors increase computational load and storage requirements, potentially leading to prolonged training times and excessive resource consumption.
3.4.2 Late Fusion-Decision Weighting: The decision weighting method in late fusion was proposed by Zadeh et al. in 2016 [1]. This approach processes speech and text separately to obtain unimodal decision results, which are then fused via weighted averaging. The advantage of this fusion strategy lies in its high flexibility, making it suitable for scenarios with significant modal discrepancies [2]. Different modalities can adopt optimized processing methods tailored to their characteristics before fusion, maximizing their individual strengths. For example, in data-to-text generation tasks, text and data can be processed independently and then fused using specialized algorithms to produce coherent outputs [8]. However, late fusion has limitations: it ignores deep interactions between modalities, potentially failing to exploit latent cross-modal relationships.
3.4.3 Intermediate Fusion–Attention Mechanisms in Intermediate Fusion: Intermediate fusion performs feature fusion at the intermediate layers of the model, with representative architectures including the Tensor Fusion Network (TFN) and Multimodal Transformer (MulT) [1,8]. This fusion approach typically employs attention mechanisms to capture cross-modal relationships at different hierarchical levels, enabling more effective integration of multimodal information. For example, in multimodal emotion recognition, attention mechanisms can fuse speech and facial expression features at intermediate layers to enhance emotion recognition accuracy [1]. By dynamically adjusting weights according to modal importance, attention mechanisms allow the model to prioritize critical information, thereby improving overall performance [15,16].

3.5 Deep Learning Model

3.5.1 LSTM and Variants: Leverage the temporal modeling capability of LSTM while incorporating attention mechanisms to capture critical emotional cues.
ICON: Models dialogue history using bidirectional GRU and captures self-interaction and mutual emotional influences via a multi-hop

memory network [1].
BiLSTM+Attention [19]: Applies BiLSTM to text and speech separately, then focuses on key frames/words through attention mechanisms [15].
3.5.2 Transformer and Attention Mechanisms: Leverage self-attention and cross-attention mechanisms to model long-range dependencies between modalities.
Multimodal Transformer:
Cross-Attention Module: Learns dynamic alignment between modalities (e.g., vision→language, language→audio)[2,8].
Self-Attention Layer: Integrates fused features for prediction.
Transformer-xl [20]: Extends Transformer to handle long sequences using relative positional encoding.
3.5.3 Graph Neural Network (GNN): Models cross-modal interactions as graph structures to capture dynamic relationships [21].
Dynamic Fusion Graph:
Nodes: Represent unimodal, bimodal, and trimodal interactions.
Edges: Dynamically adjust connection strengths via attention mechanisms[1].
Graph-MFN:
Combines memory networks[22] with graph structures to dynamically store modal interactions[1].

4. Challenges and Difficulties

Despite significant advances in deep learning-based multimodal emotion analysis in recent years, practical applications still face multiple challenges. These challenges stem from both the complexity inherent in the data and limitations in model design.

4.1 Data Alignment Issue

Temporal discrepancies between speech and text[2,23] make direct cross-modal alignment challenging, hindering feature interaction learning. Key manifestations include:
Sampling rate discrepancy: Speech signals (e.g., 16kHz) and text word-level granularity (average 0.5–1 second per word) exhibit order-of-magnitude differences.
Asynchrony: Emotional expressions may precede or lag text (e.g., a smile during the pronunciation of "好" ("good") may occur before or after the word).
Existing solutions to these issues include two approaches: explicit alignment and implicit alignment.
Explicit alignment:①Dynamic Time Warping (DTW): Computes a distance matrix between modalities to find an optimal temporal alignment path[24].②Forced alignment tools (e.g., P2FA[23]): Segment speech and text into identical time slices using phoneme alignment. However, explicit alignment methods suffer from:High computational complexity、Reliance on prior alignment assumptions、Potential loss of aligned emotional cues.
Implicit alignment: Attention mechanisms dynamically learn cross-modal dependencies via cross-attention. The formula is: $Y_\alpha = \mathrm{softmax}\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right)V_\beta$ [15], where Q_α and K_β denote queries and keys for text and speech, respectively.
Data alignment still requires reinforcement learning to dynamically adjust alignment strategies, such as optimizing attention weights via reward mechanisms. Alternatively, exploring unsupervised alignment methods using self-supervised learning (e.g., contrastive learning) can reduce reliance on alignment annotations.

4.2 Modal Complementarity Mining

The complementary information between semantic (text) and prosodic (speech) modalities is difficult to effectively fuse, particularly in complex scenarios (e.g., sarcasm, irony). Key manifestations include:
Sarcasm Detection: The text "This movie is amazing!" paired with a sneering tone indicates a negative true emotion [19].
Emotional Intensity Discrepancy: The text "not bad" may correspond to neutral or slightly negative intonation in speech.
For emotion recognition in these complex scenarios, existing approaches primarily involve two methods: attention mechanisms and adversarial learning.
Attention Mechanisms: BiLSTM+Attention [19]: Focuses on negative words (e.g., "no") in text and high-energy frames in speech via attention mechanisms.
Adversarial Learning: Adversarial Fusion: For example, Pham's MCTN model forces cross-modal complementary information maximization via adversarial training. [25,26].
Future research could design modal importance evaluation metrics to quantify complementary contributions [1] and integrate psychological theories (e.g., emotional contagion theory) to

model dynamic inter-modal interactions.

4.3 Data Scarcity and Domain Adaptation

Insufficient labeled data and high annotation costs, combined with the subjectivity of manually assigned emotional labels, lead to model overfitting and limited cross-domain generalization (e.g., from customer service dialogues to social media). Key manifestations include:

Few-Shot Scenarios: For example, high annotation costs for mental health dialogues make it difficult to cover all emotion types [1].

Domain Differences: Significant differences exist between informal language in social media (e.g., "阿弥诺斯" (a colloquial term)) and formal expressions in customer service dialogues [19].

Existing solutions to these issues include:

Pre-trained Models:

BERT: Pre-trained on large-scale text corpora and fine-tuned for emotion analysis [27].

HuBERT: Hsu [10] extracts features from unlabeled speech via self-supervised learning to improve cross-domain performance.

Few-Shot Learning:

Meta-Learning: For example, MAML (Model-Agnostic Meta-Learning) [28] quickly adapts to new domains with small amounts of labeled data.

Future research could develop multimodal self-supervised learning frameworks (e.g., contrastive learning [29] for cross-modal samples) or utilize domain adaptation techniques [30] to reduce domain discrepancies.

4.4 Model Complexity and Interpretability

Deep models (e.g., Transformers, GNNs) with large numbers of parameters and complex architectures are difficult to interpret in decision-making, limiting practical applications. Meanwhile, real-time requirements impose additional constraints. Key manifestations include:

Black-Box Problem: Doctors find it difficult to trust AI's emotional judgments of patients, requiring interpretability support [31,32].

Computational Cost: Deep networks require multi-GPU training and are unsuitable for edge devices [29].

Visualization Techniques:

Attention Heatmaps: Cross-attention weights from MulT (Multimodal Transformer) [8] localize key emotional cues (e.g., the association between "disappointing" and angry facial expressions).

LIME/SHAP: Local explanations of model decisions, such as how negative words in text influence emotion predictions [31,32].

Model Compression: Knowledge Distillation [33]: Transfers knowledge from complex models (e.g., Transformers) to lightweight models (e.g., TinyBERT).

To address this challenge, future research could develop interpretability frameworks, such as integrating logical rules with neural networks or exploring dynamic computation models that adjust computational load based on input complexity [2,3].

5. Conclusion

Multimodal emotion analysis based on speech and text significantly enhances accuracy and robustness by fusing semantic and prosodic information. This paper systematically reviews core methods and challenges in this field: from early shallow fusion via feature concatenation to deep interactions using attention mechanisms and graph neural networks [1,19]; from discrete emotion classification to continuous-dimensional modeling, technological evolution has consistently revolved around modal complementarity. However, challenges such as data alignment difficulties, insufficient generalization in complex scenarios, and poor model interpretability remain unresolved. Future advancements in multimodal emotion analysis must break through in four dimensions: efficiency, generalization, personalization, and dynamism. Efficient fusion methods reduce deployment costs, cross-modal transfer learning addresses data scarcity, personalized analysis improves user experience, and multimodal dialogue systems drive practical applications. Combining self-supervised learning, lightweight models, and dynamic graph structures will facilitate the widespread adoption of emotion analysis technology in domains such as intelligent customer service and mental health monitoring.

References

[1] A. Zadeh, R. Zellers, E. Pincus and L. -P. Morency, "Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages," in IEEE Intelligent Systems, vol. 31, no. 6, pp. 82-88, Nov.-Dec. 2016, doi: 10.1109/MIS.2016.94.

[2] T. Baltrušaitis, C. Ahuja and L. -P. Morency,

"Multimodal Machine Learning: A Survey and Taxonomy," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[3] TAO Jianhua, FAN Cunhang, LIAN Zheng, et al. (2024). Current developments and trends in multimodal emotion recognition and understanding. Journal of Image and Graphics, 29 (6), 1607–1627. https://doi.org/10.11834/jig.240017

[4] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

[5] Busso C , Bulut M , Lee C C ,et al.IEMOCAP: interactive emotional dyadic motion capture database[J].Language Resources and Evaluation, 2008, 42(4):335-359.DOI:10.1007/s10579-008-9076-6.

[6] Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of multimodal sentiment recognition and understanding. Journal of Image and Graphics, 29 (06:1607-1627

[7] Tassi, D., Marchi, E., & Zaccaria, R. (2015). Speech emotion recognition using MFCC features and HMM classifiers. Procedia Computer Science, 62, 1152-1159.

[8] Trigeorgis, G., Ringeval, F., Brueckner, R., Maciejewski, M., & Schuller, B. W. (2016). AdieuNet: Deep audio-visual emotion recognition with multiplicative LSTM integration. In 2016 16th IEEE International Conference on Data Mining Workshops (ICDMW), pp. 57–64.

[9] Wollmer, M., Eyben, F., Schuller, B. W., & Rigoll, G. (2013). Recurrent neural networks for robust speech emotion recognition. In 2013 21st European Signal Processing Conference (EUSIPCO), pp. 1-5.

[10] Hsu, W. -N., Tang, Y., Hsiao, Y., Chuang, Y. -T., Lee, S. -W., & Lin, S. -W. (2021). HuBERT: Self-supervised learning for spoken language understanding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5835-5846.

[11] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543.

[12] Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media texts. In Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM), pp. 216-225.

[13] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[15] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR).

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NIPS 2017), pp. 5998-6008.

[17] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

[18] McKeown, G., Valstar, M. F., Cowie, R., & Pantic, M. (2012). SEMAINE: A multimodal database of emotionally coloured conversations. In 2012 5th International Workshop on Database and Expert Systems Applications (DEXA), pp. 432-436.

[19] Poria, S., Cambria, E., Hazarika, D., &

Hussain, A. (2016). A review of deep learning techniques for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 7(4), 377-393.
[20] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860.
[21] Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. IEEE Transactions on Neural Networks, 20 (1), 61-80.
[22] Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. arXiv preprint arXiv:1410.3916.
[23] Yuan, Y., & Liberman, M. (2008). P2FA: A tool for forced alignment of speech and text. Language Resources and Evaluation, 42 (4), 359-376.
[24] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding and V. Tarokh, "Speech Emotion Recognition with Dual-Sequence LSTM Architecture," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6474-6478,
[25] Pham, H. L., Nguyen, T. N., & Phung, D. Q. (2019). Multimodal context transformation network for sentiment analysis. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 503-512.
[26] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. Advances in Neural Information Processing Systems (NIPS 2014), 2672-2680.
[27] Devlin, J., Choe, M. -W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 4171-4186.
[28] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 1126-1135.
[29] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748-8763.
[30] Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (7), 1418-1432.*
[31] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.
[32] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS 2017, 4765-4774.
[33] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.