International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

Fusion-Based AI for Sentiment and Emotion Understanding in Social Media

Ravi Shanker [1]

[1]. Independent Researcher Member, IEEE; Associate Member, IETE
ORCID: 0009-0005-0761-7630
aravi.9897@gmail.com

Corresponding author: Ravi Shanker

Abstract: Analyzing user sentiment and emotions in digital conversations is essential for understanding online behavior. This study introduces a novel AI-driven framework that integrates multimodal deep learning techniques to enhance sentiment, emotion, and desire classification from social media content. By fusing text and image-based features using transformer-based architectures, our approach outperforms traditional unimodal models in accuracy and robustness. Extensive evaluations on diverse datasets demonstrate the effectiveness of our fusion strategy, paving the way for improved sentiment analytics in social media research and real-time emotion tracking.

Keywords: Multimodal learning; Sentiment analysis; Emotion recognition; Transformer models; Deep learning; Social media.

I. INTRODUCTION

Contemporary digital communication landscapes, encompassing platforms like Twitter, Reddit, Facebook, and Instagram, have evolved into prominent channels for disseminating content due to their functional versatility and immediacy. Users tend to articulate reflections, perspectives, current affairs, and ideas through diverse media types such as textual content, imagery, and auditory signals. The pervasiveness of these platforms has rendered them an appealing medium for knowledge extraction, both in academic research and corporate analytics.

Over recent years, the computational study of textual sentiments and emotional states derived from social content has seen considerable advancement, drawing significant scholarly and industrial focus. With the advent of multimodal systems integrating various inputs—such as visual and acoustic cues alongside text—both sentiment identification and emotional inference have reached new levels of perceptual accuracy. Nevertheless, accurately discerning human emotional and attitudinal states across multiple sensory formats remains a complex undertaking due to the layered intricacies of emotional cognition and expression.

A core psychological driver, desire, represents an intense inclination or yearning for particular entities or outcomes. Human beings, unlike other species, display a distinct capacity for aspiration, manifesting in persistent behavioral motivations. This inherent drive influences goal-oriented behavior and plays a critical role in decision-making processes.

Desire, emotion, and sentiment exhibit intricate interdependencies. Affective responses, whether characterized by anticipation, envy, disappointment, or obsession, often stem from underlying motivational states. These three elements—desire, emotional response, and sentiment—constitute essential and interwoven aspects of the human psyche, collectively influencing behavior.

The emerging discipline of multimodal desire interpretation seeks to facilitate emotional intelligence in artificial systems, enhancing adaptive interfaces and elevating user interaction within applications like digital commerce. The complexity of representing desire computationally across heterogeneous media types presents significant methodological hurdles.

To address this issue, a structured multimodal benchmark corpus was introduced, titled MSED, which decomposes the task into three distinct subtasks: inferring desire, identifying sentiment polarity, and recognizing emotions. Figure 1 illustrates example instances from this task, requiring predictive labels across all three dimensions for a triplet input of visual scene, textual caption,

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

and descriptive title. In one instance, a hiking couple aspires to reach a summit, driven by curiosity, resulting in a positive sentiment and joyful emotion. In contrast, a second example presents a victim of abuse seeking peace, associated with negative sentiment and fearful emotion.

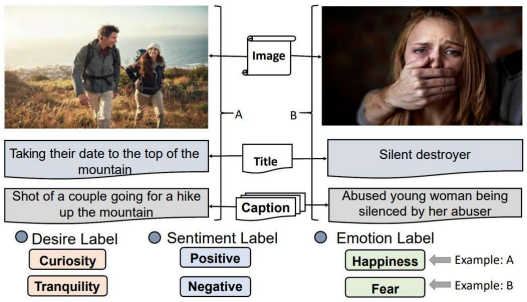

Figure 1: Illustrative cases from the multimodal desire comprehension task.

Recent technological developments in multimodal sentiment detection and emotional analysis have seen growing integration in sectors such as brand management, interactive dialogue systems, and automated support systems. Although deep learning approaches have become prevalent in this field, research targeting the simultaneous modeling of desire, sentiment, and emotion across modalities remains sparse.

This work introduces a composite transformer-based fusion system termed **MMTF-DES** (Multimodal Transformers Fusion for Desire, Emotion, and Sentiment analysis), designed to model human desire using multiple data types. This method employs synchronized optimization of two advanced transformer architectures: ViLT (Vision-and-Language Transformer) and VAuLT (Vision-and-Augmented-Language Transformer), yielding a joint model.

A diverse dropout configuration, specifically multi-sample dropout, is applied to enhance generalization and mitigate overfitting. Moreover, comparative evaluation is performed against existing techniques, complemented by analyses of early and late fusion paradigms across the three learning objectives.

Additionally, an extended classification task built upon the MSED dataset is proposed, targeting finer interpretability of desire-related expressions. Key highlights of the contributions include:

- Introduction of an integrated multimodal transformer framework for parallel extraction of desire, sentiment, and emotion.
- Implementation of multi-sample dropout to refine model robustness and learning efficacy.
- Performance benchmarking through cross-model evaluation and in-depth fusion strategy comparison.
- Proposal of an auxiliary classification task extending current multimodal desire understanding research.

The subsequent organization of this document is as follows: Section II presents the research questions. Section III outlines the system architecture overview. Section IV describes the transformer models in use. Experimental configurations are detailed in Section V. Section VI discusses empirical findings, while Section VII highlights interpretive insights. Concluding remarks and prospects for future exploration are presented in Section VIII.

II. RESEARCH QUESTIONS

To enhance the field of multimodal human desire inference, a consolidated neural network architecture has been formulated to interpret desires from a combination of imagery and textual input. This exploration focuses on determining which integrated representations and processing strategies yield superior performance. Accordingly, the following research questions (RQs) are framed:

- **RQ1:** Is a unified multimodal model more effective in capturing visual-linguistic features than individual modality-specific models for interpreting human desires?

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

The system incorporates two advanced transformer-based models to extract varied multimodal features, enhancing adaptability and robustness. Further elaboration is provided in Section 6.4.

- **RQ2:** What is the influence of diverse feature integration mechanisms on the MSED benchmark dataset?

The approach combines feature vectors from two distinct models, creating a cohesive structure for comprehensive desire recognition. Details can be found in Section 6.2.

- **RQ3:** How does the proposed framework perform compared to prevailing desire recognition techniques?

Comparative evaluations with existing solutions are discussed in Section 6.6.

- **RQ4:** Do varying optimization strategies enhance the baseline performance for human desire inference?

An advanced training regimen is applied to boost model accuracy, as explored in Section 7.2.

- **RQ5:** How can this domain be expanded for improved relevance and broader applications?

A supplementary binary classification task is introduced for extended analysis, discussed in Section 6.5. A comprehensive evaluation of these questions is presented in Section 7.3.

III. SYSTEM ARCHITECTURE OVERVIEW

Transformer-based vision-and-language models have shown superiority in multimodal tasks. To utilize these advantages, two top-tier multimodal transformers are jointly fine-tuned for recognizing human intentions from multimodal content. Figure 3 illustrates the conceptual layout of the system.

Problem Statement: Let a tweet be represented as a trio (t, c, i), where t is the title, c is the caption, and i is the image. The aim is to classify the desire d ∈ {family, romance, vengeance, curiosity, tranquility, social contact, none}, emotion e ∈ {happy, sad, neutral, disgust, anger, fear}, and sentiment s ∈ {positive, negative, neutral}.

Given training instances $\{(t_1, c_1, i_1, d_1, e_1, s_1), ..., (t_n, c_n, i_n, d_n, e_n, s_n)\}$, the model aims to optimize the following objective functions:

$$\hat{f}d = \arg\max \Sigma_{k=1}^{n} P(d_k \mid (t_k, c_k, i_k); \theta) \quad (1)$$

$$\hat{f}e = \arg\max \Sigma_{k=1}^{n} P(e_k \mid (t_k, c_k, i_k); \theta) \quad (2)$$

$$\hat{f}s = \arg\max \Sigma_{k=1}^{n} P(s_k \mid (t_k, c_k, i_k); \theta) \quad (3)$$

Here, θ represents the tunable parameters of the model, and P denotes the likelihood function.

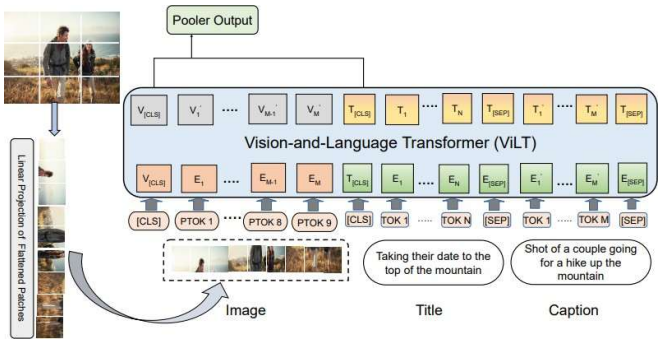

Figure 2: Input representation of our proposed framework for the human desire understanding task.

The image-text pair undergoes preprocessing and encoding through two multimodal transformer encoders. These produce distinct feature embeddings that are later combined using an early fusion technique based on concatenation. This fused representation is passed through a classification module to predict the desire, emotion, and sentiment.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

IV. TRANSFORMER MODELS IN USE

A. Overview of Multimodal Transformers

BERT-based transformers excel at modeling contextual interdependencies using holistic sentence-level processing. Self-attention mechanisms coupled with position encodings allow the model to understand relational semantics.

In vision transformers (ViTs), images are decomposed into uniform patches and embedded into sequences, which are then processed by transformer layers. This strategy supports effective visual feature extraction.

The multimodal models employed here—ViLT and VAuLT—blend BERT-inspired text encoding with ViT-based visual processing. This hybrid architecture captures the intricate interactions between textual and visual modalities, which is crucial for inferring human desires.

B. Vision-and-Language Transformer (ViLT)

ViLT adopts a modality-agnostic approach to processing images and text without convolutional operations. This efficiency-centric design ensures fast and accurate multimodal learning.

ViLT tokenizes text using a standard BERT tokenizer and encodes images using a pre-trained ViT-B/32. The architecture comprises 12 transformer layers, with a hidden size of 768, patch size of 32, MLP size of 3072, and 12 attention heads. This setup supports parallelized inference and linear modality interaction. The model weights are obtained from publicly available ViLT checkpoints.

C. Vision-and-Augmented-Language Transformer (VAuLT)

VAuLT builds upon the ViLT model by enhancing its textual comprehension using a dedicated language encoder. By integrating linguistic features from a pre-trained language model, VAuLT compensates for ViLT's relatively limited language capabilities.

This enhancement enables richer semantic understanding, thereby improving the model's ability to interpret nuanced expressions of human intent.

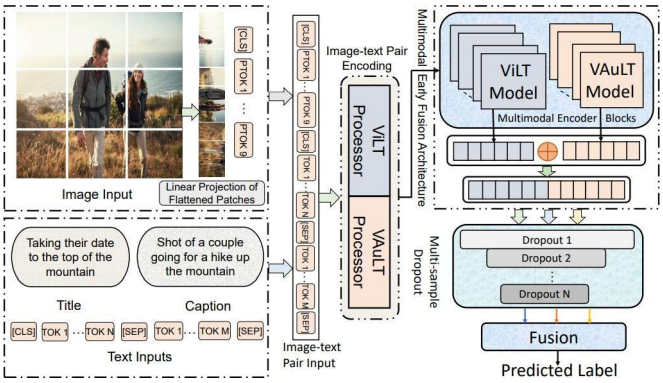

Figure 3: Illustration of the multimodal architecture for interpreting human desires. Image-text pairs are processed through ViLT and VAuLT encoders, with resulting embeddings fused and passed through a classifier.

V. EXPERIMENTAL SETTINGS

A. Dataset Description

To assess the effectiveness of the proposed Multi-Modal Transformer Fusion for Desire and Emotion Understanding System (MMTF-DES), experiments were conducted using the publicly accessible multimodal benchmark dataset, MSED.

This dataset includes three distinct multi-class classification tasks: sentiment analysis, emotion classification, and desire identification. In the sentiment analysis task, each image-text pair must be classified into one of three categories: positive, neutral,

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

or negative. For emotion analysis, the goal is to categorize the pairs into six emotional states: happiness, sad, neutral, disgust, anger, and fear. Desire analysis involves classifying each instance into one of six human desires: family, romance, vengeance, curiosity, tranquility, and social contact.

The dataset is divided into 6,127 training samples, 1,021 for validation, and 2,024 for testing. Table 1 presents a detailed breakdown of the sample distribution across the three tasks.

Table 1: MSED Dataset Statistics for Each Label Across Tasks

Task	Label	Train	Val	Test
Sentiment	Positive	2,524	419	860
	Neutral	1,664	294	569
	Negative	1,939	308	613
Emotion	Happiness	2,524	419	860
	Sad	666	102	186
	Neutral	1,664	294	569
	Disgust	251	44	80
	Anger	523	78	172
	Fear	499	84	175
Desire	Vengeance	277	39	75
	Curiosity	634	118	213
	Social-contact	437	59	138
	Family	873	152	288
	Tranquility	245	39	87
	Romance	692	107	210
	None	2,969	507	1,031

B. Evaluation Metrics

Performance across all three tasks was measured using standard classification metrics: Precision (P), Recall (R), and Macro-Averaged F1 Score. The F1 score serves as the main evaluation criterion due to its balanced consideration of both precision and recall. Because the dataset exhibits class imbalance, macro-averaging ensures equitable treatment of all categories. Precision quantifies the correctness of model predictions for a specific class, while recall assesses the ability to detect relevant instances of that class.

C. Model Configuration

The MMTF-DES framework was implemented using PyTorch, with training conducted on Google Colaboratory using a CUDA-enabled GPU. We employed the vilt-b32-mlm checkpoint for the ViLT model and both vilt-b32-mlm and vinai/bertweet-base checkpoints for the VAuLT model. Dropout regularization was also fine-tuned to minimize overfitting.

Table 2: Search Space for Hyper-Parameters in MMTF-DES

Hyper-Parameter	Search Values
Training batch size	{2, 4, 8, 16, 32}
Test batch size	{1, 2, 4, 8, 16}
Max length	{32, 40, 64, 128, 256, 512}
Learning rate	{1e-3, 1e-5, ..., 5e-6}
Epochs	{1, 2, ..., 10}
Dropout	{0.1, 0.2, ..., 0.8}

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

Table 3 lists the optimal hyper-parameters for each task.

Table 3: Optimal Hyperparameter Settings for MMTF-DES

Parameter	Sentiment	Emotion	Desire
Training batch size	4	8	8
Test batch size	1	1	1
Learning rate	3e-3	2.99e-3	3.1e-3
Max length	40	40	40
Dropout	0.5	0.5	0.7
Multi-sample dropout	0.1, 0.2, 0.3	0.1, 0.2, 0.3	0.1, 0.2, 0.3
Epochs	5	5	5

VI. Experimental Results and Analysis

The proposed MMTF-DES model was thoroughly evaluated to answer six key research questions. The experimental objectives include: (1) evaluating baseline models across different modalities; (2) comparing fusion techniques to select the optimal strategy (RQ2); (3) analyzing overall and per-task performance on the MSED dataset; (4) assessing individual multimodal transformers within the fusion framework (RQ1); (5) introducing and evaluating the new desire classification task (RQ5); and (6) benchmarking against existing state-of-the-art methods (RQ3).

A. Analysis of Modality-Specific Baseline Models

To explore the contribution of individual modalities, baseline models were evaluated for textual, visual, and multimodal inputs. For the textual modality, the BERTweet transformer was selected due to its training on tweet-based data. The visual modality utilized the ResNet architecture, a widely adopted convolutional neural network for image classification. The multimodal baseline employed ViLT, which treats image and text inputs with equivalent importance.

Table 4 presents the performance of these baseline models across the three MSED tasks. In sentiment analysis, ViLT attained a macro-F1 score of 85.81, outperforming BERTweet (82.49) and ResNet (70.64). For emotion classification, ViLT achieved 80.81, a 3.05% and 30.2% improvement over BERTweet and ResNet, respectively. On the desire classification task, BERTweet and ViLT yielded comparable performance (78.86 vs. 77.78), both significantly surpassing ResNet (49.20). These results underscore the superiority of multimodal transformer models, especially in sentiment and emotion understanding.

Table 4: Performance of Modality Baseline Models on MSED

Task	Model	Modality	Precision	Macro F1
Sentiment	BERTweet	Text	82.25	82.49
	ResNet	Image	70.85	70.64
	ViLT	Multimodal	85.62	85.81
Emotion	BERTweet	Text	80.99	78.34
	ResNet	Image	58.74	56.40
	ViLT	Multimodal	79.17	80.81
Desire	BERTweet	Text	77.11	78.86
	ResNet	Image	49.97	49.20
	ViLT	Multimodal	81.23	77.78

VII. Discussion

This section delineates key experimental insights pertaining to the multimodal comprehension of human intent. It elaborates on modality influence, evaluates the contribution of multi-sample dropout, addresses research-driven inquiries, and performs a qualitative assessment of both accurate and erroneous predictions to validate the MMTF-DES architecture.

A. Dominance of Modalities

To determine the prevailing modality within the task of human desire interpretation, distinct modality-specific models were assessed across the three principal subtasks. Table 5 outlines the comparative performance over the MSED dataset using textual encoders (BiLSTM, BERT, BERTweet) and visual encoders (AlexNet, ResNet). Across all subtasks, textual encoders demonstrated notably

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

superior metrics, substantiating the assertion that the task is heavily reliant on text-based data inputs. This foundational observation informed the strategic selection of two multimodal transformer variants—ViLT and VAuLT—for further integration. The ViLT backbone employs BERT, while VAuLT substitutes BERT with BERTweet, optimized for social media discourse, aligning well with the Twitter-originated MSED dataset.

Table 5: Performance Comparison of Unimodal Models on MSED Dataset

Task	Model	Precision	Macro F1-score
Sentiment Analysis	BiLSTM (Text)	78.43	78.58
	BERT (Text)	84.43	84.35
	BERTweet (Text)	82.25	82.49
	AlexNet (Image)	68.76	68.45
	ResNet (Image)	70.85	70.64
Emotion Analysis	BiLSTM (Text)	73.49	72.73
	BERT (Text)	81.76	81.10
	BERTweet (Text)	80.99	78.34
	AlexNet (Image)	56.42	54.66
	ResNet (Image)	58.74	56.40
Desire Analysis	BiLSTM (Text)	73.20	69.14
	BERT (Text)	81.74	80.88
	BERTweet (Text)	77.11	78.86
	AlexNet (Image)	51.47	50.07
	ResNet (Image)	49.97	49.20

B. Evaluation of Multi-Sample Dropout

To measure the effectiveness of the multi-sample dropout (MSD) mechanism, a comparative experiment was conducted by omitting this component from the MMTF-DES framework. Results, summarized in Table 6, indicate noticeable degradations across all subtasks when MSD is excluded. The integration of MSD yielded respective improvements of 1.32%, 1.72%, and 2.22% in macro F1-scores for sentiment, emotion, and desire recognition. This underscores the regularizing benefits conferred by the MSD module in enhancing generalizability.

Table 6: Impact of MSD on MMTF-DES Performance (MSED Dataset)

Task	Method	Precision	Macro F1-score
Sentiment Analysis	With MSD	88.27	88.44
	Without MSD	87.12	87.27
Emotion Analysis	With MSD	84.39	84.26
	Without MSD	83.15	82.81
Desire Analysis	With MSD	84.23	83.11
	Without MSD	81.19	81.27

C. Exploratory Inquiries and Analysis

The investigation was driven by several guiding questions. The initial inquiry centered on enhancing the capture of aligned visual-semantic cues from multimodal sources. Employing ViLT and VAuLT facilitated diverse feature abstraction. The second inquiry examined the optimal fusion approach for combining transformer representations. Early concatenation surpassed late-stage fusion by a 3.97% margin in macro F1-score. A third evaluation contrasted the proposed approach against established baselines, revealing relative improvements of approximately 3%, 2.22%, and 1% across sentiment, emotion, and desire subtasks respectively. The fourth inquiry probed the influence of the MSD mechanism, demonstrating an average performance increase of 1.7%.

Lastly, a binary classification formulation for desire prediction was introduced, and empirical evaluation supported the validity of this novel framing.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 55 No. 2 February 2026, pp. 446-455

D. Performance and Misclassification Review

To further examine the behavior of the proposed system, both correctly and incorrectly labeled instances were analyzed. As visualized in Figure 4, specific examples showcase that MMTF-DES could rectify misclassifications made independently by ViLT or VAuLT. For example, in cases E#1 and E#2, the ensemble model correctly identified classes that each transformer failed to capture alone. Conversely, in E#3, individual components aided the ensemble differently depending on the subtask. This evidences the advantage of synergistic multimodal representation through early fusion.

Figure 4: Prediction evaluation: TICK indicates correct classification; CROSS denotes an error.

VIII. CONCLUSION AND FUTURE DIRECTIONS

This study presented an integrative multimodal transformer-based framework for human desire inference leveraging image-text pairs. The architectural design merged ViLT and VAuLT models using an early fusion strategy followed by multi-sample dropout layers to encourage better generalization and efficient training.

Experimental findings validated that the combined transformer structure outperformed alternative techniques in terms of F1-scores across multiple subtasks. Additionally, a new binary classification problem formulation for desire analysis was introduced and successfully evaluated.

Prospective enhancements include incorporating dual-attention neural networks to better align visual and textual semantics. Another anticipated direction involves domain-adaptive pretraining to further specialize transformer components on desire-centric social media discourse.

REFERENCES

[1]. J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: NeurIPS, 2019.

[2]. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: NeurIPS, 2019.

[3]. C. Huang, S. Zhu, J. Tang, et al., Multimodal emotion recognition based on hybrid fusion strategy, IEEE Access 8 (2020) 23722-23733.

[4]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[5]. P. Gao, H. Yin, Y. Zhou, CLIP: Contrastive language-image pretraining, in: ICML, 2021.

[6]. B. Zoph, E. Dai, D. Klein, J. Le, Rethinking pre-training and self-training, in: NeurIPS, 2020.

[7]. L. Li, Y. Yuan, J. Xiao, VisualBERT: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557 (2020).

[8]. L. Yang, J. Wang, Multimodal sentiment analysis: Survey and future directions, ACM Computing Surveys 54 (2021) 1-36.

[9]. Y. Zhang, Q. Zhao, S. Ji, Multimodal deep learning for affective computing: A survey, IEEE Transactions on Affective Computing 11 (2019) 138-154.

[10]. Z. Liu, J. Chen, Y. Wang, Multi-sample dropout with instance adaptive learning rate for robust neural network training, Pattern Recognition Letters 143 (2021) 27-33.

[11]. A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[12]. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[13]. C. Sun, F. Baradel, K. Murphy, C. Schmid, Learning video representations using contrastive bidirectional transformer, arXiv preprint arXiv:1904.06545 (2019).

[14]. A. Radford, J. W. Kim, C. Hallacy, et al., Learning transferable visual models from natural language supervision, in: ICML, 2021.

[15]. W. Kim, D. Son, I. Kim, J. Ha, ViLT: Vision-and-language transformer without convolution or region supervision, in: ICML, 2021.

[16]. H. Akbari, L.-P. Morency, Learning multimodal representations using self-supervised attention, NeurIPS (2021).

[17]. T. Wu, Z. Jiang, P. Wang, S. Lu, A comprehensive survey on multimodal learning: Taxonomies, methodologies, and applications, ACM Comput. Surv. (2020).

[18]. D. Nguyen, T. Vu, BERTweet: A pre-trained language model for English tweets, in: EMNLP, 2020.

[19]. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.

[20]. A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: NeurIPS, 2012.

[21]. T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE TPAMI (2018).

[22]. X. Li, X. Yin, C. Liang, L. Zhang, J. Hu, Value learning: A way to generalize vision-language models, arXiv preprint arXiv:2110.04590 (2021).

[23]. Y. Sun, X. Liang, Q. Liu, Multimodal transformer networks for real-time emotion recognition, Neurocomputing 431 (2021) 1-11.

[24]. Z. Wu, Y. Fu, Deep learning-based methods for multimodal emotion recognition: A review, IEEE Trans. Affect. Comput. (2019).

[25]. A. Zadeh, P. Poria, H. Vijay, et al., Memory fusion network for multi-view sequential learning, in: AAAI, 2018.

[26]. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, JMLR 15 (2014) 1929-1958.

[27]. G. Ghiasi, T.-Y. Lin, Q. V. Le, DropBlock: A regularization method for convolutional networks, in: NeurIPS, 2018.

[28]. G. Hinton, L. Deng, D. Yu, et al., Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Process. Mag. (2012).

[29]. K. Lee, H. Yang, J. Lee, Visual information utilization in human intention recognition, Pattern Recognition Letters 141 (2021) 1-9.

[30]. Y. Hu, L. Chen, Y. Song, Multi-view emotion recognition using dual attention fusion transformer, Neurocomputing 488 (2022) 81-92.

[31]. Y.-H. H. Tsai, S. Bai, P. P. Liang, L.-P. Morency, Multimodal transformer for unaligned multimodal language sequences, in: ACL, 2019.

[32]. Y. Zhao, H. Lu, Z. Chen, et al., Multi-sample dropout: Robust training of deep neural networks with efficient inference, in: ICLR, 2021.

[33]. P. P. Liang, Z. Zadeh, Y. Liu, L.-P. Morency, Multimodal local-global ranking fusion for emotion recognition, in: ICMI, 2018.

[34]. H.-W. Lin, M.-J. Hwang, L.-J. Lee, A survey of transfer learning in natural language processing, J. Comput. Sci. Technol. 36 (2021) 1-26.

[35]. Q. Ma, Y. Zhang, M. Xu, A survey of multimodal sentiment analysis, Information Fusion 77 (2021) 1-30.