Multimodal Sentiment Analysis Based on Fused Momentum Distillation and Contrastive Learning

GAN Quan[*], WANG Chen[*], WU Jiaying[*], ZHONG Zhaoman[*], GE Qi-Wei[**], LIU Chuanxia[***, †]

（Abstract）

With the rapid development of social media, the nature of the data it generates has evolved from single-text to multimodal formats. This evolution has given rise to the multimodal sentiment analysis task, which aims to comprehensively analyze information from multiple modalities, such as text and images. However, the inherent heterogeneity of multimodal data presents significant challenges. To overcome these challenges, we propose a multimodal data fusion model that integrates momentum distillation and contrastive learning to enhance the alignment and fusion of these heterogeneous modalities. Our approach first utilizes two single-modal encoders to obtain representations for text and image inputs, which are then integrated through attention-based modal fusion. These fused representations are subsequently fed into a momentum distillation model to construct negative samples for image-text matching and image-text contrast training, thereby facilitating modality alignment and improving the learning of correlations between different modalities. Finally, experiments conducted on two public datasets (“MVSA-single” and “MVSA-multi”) demonstrate the superiority of our model in multimodal sentiment analysis, highlighting the contributions of the proposed modules in enhancing analysis effectiveness.

Keywords: multimodal sentiment analysis, heterogeneity of multimodal data, momentum distillation, contrastive learning

1. Introducation

The rapid development of social media and mobile communication devices over the past few decades has dramatically changed the way people express their emotions. With the development of increasingly ubiquitous networks and more versatile mobile devices, people are increasingly using a combination of images and text to express more complex and rich emotions. In this new form of expression, it is often difficult to accurately capture people's emotional tendencies only by analyzing the text, because the non-verbal information contained in images plays a crucial role in emotional expression. Therefore, a multimodal data processing method is needed for sentiment analysis.

Multimodal data encompasses various types of information, including visual, auditory, speech, and text data. For instance, within a video segment, there could be visual details regarding color, size, and shape, along with audio components like BGM. To better capture the author's emotions, a comprehensive consideration and analysis of the information embedded in multimodal data is necessary. Multimodal Sentiment Analysis (MSA) is leveraging information from multiple modalities to infer or understand the emotional state [1]. This approach broadens the scope of traditional sentiment analysis, enhancing its adaptability to the diverse sensory inputs encountered in the real world.

In the early research on MSA, methods such as early fusion, late fusion, and hybrid fusion were commonly employed [2]. However, in recent years, with the advancement of deep learning, the trend in research leans towards utilizing deep learning models for MSA. The reason lies in the fact that deep learning models are more flexible,

[*] School of Computer Engineer, Jiangsu Ocean University
[**] The Graduate School of East Asian Studies, Yamaguchi University
[***] School of Foreign Languages, Jiangsu Ocean University
[†] Corresponding Author

Journal of East Asian Identities Vol. 10 March 2025　（*pp. 88-95*）

adaptive, and capable of learning more complex inter-modal relationships from the data. Researchers have introduced various MSA techniques, including attention mechanisms, tensor fusion [3], contrastive learning [4], etc., aiming to further enhance performance and efficiency. Although these methods have shown success in modal fusion tasks, most of the work is based on modal fusion methods without emphasizing the heterogeneity between modalities. Heterogeneity is reflected in the fact that data in different modalities are different, e. g., the “text data” is a discrete symbolic representation, while the “images data” are continuous pixel representations. Due to the differences in their representations, the information in different modalities exists in different feature spaces, and the dimensionality, distribution, and semantics of the data are also different. The semantics of different modal data are vastly different, making it difficult to accurately capture the complex relationships between modalities during feature fusion. Effectively aligning and fusing the information of these heterogeneous modalities to obtain more accurate sentiment understanding is the key to realizing efficient multimodal analysis.

To address this issue, we propose a method using Momentum Distillation for Contrastive Learning (MDCL). In the feature extraction stage, this method transforms multimodal data into features of the same shape and employs a contrastive loss function for alignment, thereby improving the heterogeneity between modalities. The application of contrastive learning in MSA enables the model to better comprehend and align the similarities and differences among different modalities. Simultaneously, momentum distillation provides an effective knowledge transfer mechanism for the model. By using pseudo targets generated by the delayed update distillation model for training, the performance of the model is improved when processing heterogeneous modal data. Finally, by training and testing the MVSA-S and MVSA-M dataset, the method we proposed achieves better performance compared to several baseline models in all two datasets.

2. Related Work

2.1 Multimodal Sentiment analysis

The MSA task was first proposed by Morency et al. [5], early MSA focused on modality fusion methods, for pre-fusion, Convolutional Neural Networks (CNNs)[6], Long short-term memory(LSTM)[7] or Deep Neural Networks (DNNs) are commonly used to extract the unimodal features for splicing and then obtaining the sentiment polarity through the fully connected layer. As well as tensor-based fusion methods to fuse different modalities. For modal fusion based on attention mechanism, YANG et al. [8] proposed a new multi-modal sentiment analysis model, which introduced CNN and CBAM attention mechanisms after concatenating text features and image features, and Zadeh et al. [3] designed a tensor fusion network (TFN) that can gradually fuse multimodal information. For late fusion, Jiang et al. [9] proposed a fine-grained attention mechanism to interactively learn cross-modal fusion representations of visual and textual information. Due to the lack of systematic research on the degree of cross-modal feature matching at the affective-semantic level, Chen et al. [10] proposed a joint SA multimodal adaptive approach based on graphic relevance. Tasi et al. [11] proposed a directed paired cross-modal attention approach that focuses on the interactions between multimodal sequences.

2.2 Contrastive Learning

Contrastive learning is a newly emerging self-supervised learning method. It learns a good representation of the data by comparing positive and negative sample pairs, and the basic concept involves bringing anchor points and positive samples closer together while pushing away anchor points and negative samples. Contrastive learning was introduced to the MSA task to reduce the modal gap in MSA. Mai et al. [12] proposed a hybrid contrastive learning framework, which enables the model to fully explore cross-modal interactions, preserve interclass relationships, and reduce inter-modal gaps through within/between-modal contrastive learning and semi-contrastive learning. Similarly, Lin et al. [13] proposed a novel hierarchical graph contrast learning framework that first constructs unimodal graphs and then integrates these unimodal graphs to form multimodal graphs for both intramodal and intermodal graph contrast learning. Zolfaghari et al. [14] considered intramodal similarity to efficiently avoid mapping the same content to multiple points in the embedding space, solved previous loss function limitations, and defined the set of highly correlated samples to exclude them from negative samples to avoid false negative samples.

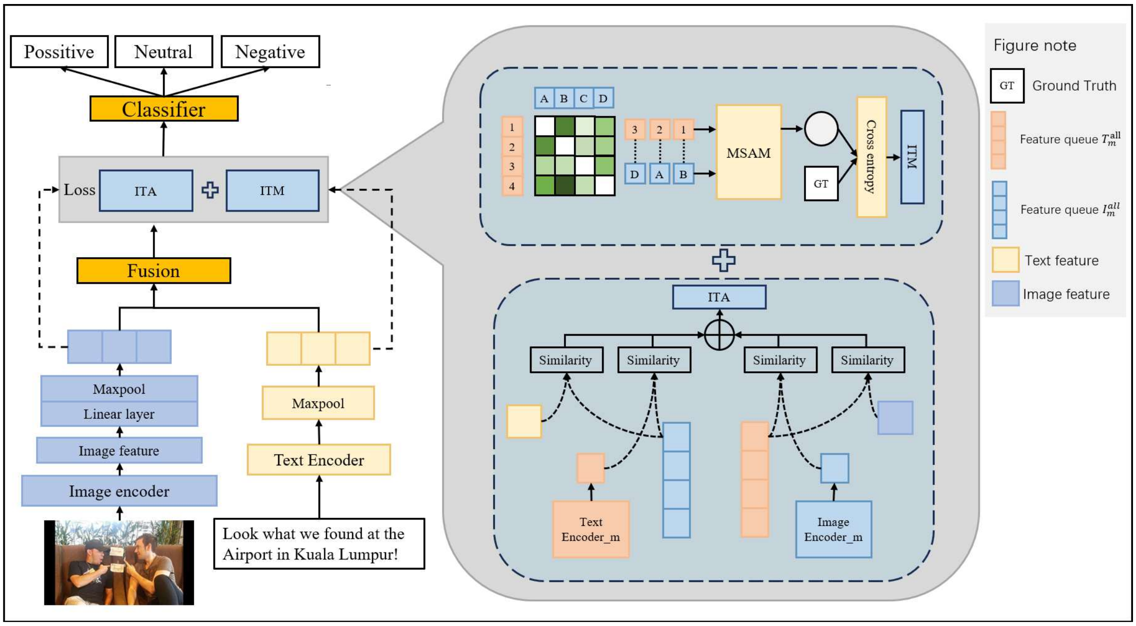

Figure 1: The framework of the MDCL model for multimodal sentiment analysis. The left side of the image are the text encoding process and the image encoding process. The fused multimodal data will pass through the classifier in the upper left corner to obtain emotion classification output. The lower half of the right part of the image is the ITA-Loss module, and the upper half is the ITM-Loss module.

2.3 Momentum Distillation

Knowledge Distillation (KD) is a model compression technique that works by transferring knowledge from a large, complex model (often called a teacher model) to a smaller, more easily deployable model (called a student model). Knowledge distillation is categorized as offline distillation, online distillation, and self-distillation. Offline distillation trains a high-performance teacher model and then uses that teacher model to guide the training process of the student model. Online distillation is a distillation method that takes place in real time during model training. It usually involves the simultaneous training of multiple student models or a student model and a teacher model, whereas self-distillation, which we have chosen to use in this paper, is where the model refines and optimizes itself without an external teacher model, and where the self-distillation uses some form of its own model as a teacher to instruct an alternative version of the model. This feature makes self-distillation much less computationally expensive. The concept of knowledge distillation was first introduced in [15] to transfer knowledge by minimizing the Kullback-Leibler divergence (KL) between the predictive logics of teachers and students. Subsequently, various KD methods [16, 17] were proposed based on [15] and further extended to distillation between intermediate features [18, 19].

3. Construction of MDCL model

In Section 2, we introduced the key technologies and the rationale behind their selection. In this paper, we build our models based on the benefits and effects of these three approaches. First, we introduce an attention mechanism to more effectively capture key features across modals and contextual information within modals. Next, we apply contrastive learning to narrow the positive sample features while pushing away the negative sample features, thereby reducing heterogeneity among the modals. Finally, we incorporate momentum distillation, leveraging the teacher-student alignment strategy to enhance the stability and consistency of the multimodal feature representation. By integrating these three approaches, we successfully improve the robustness, generalization, and accuracy of emotion recognition. This section discusses the structure and loss function design of the MDCL model.

3.1 Model architecture

In order to address the heterogeneity between modalities through modal alignment during the training phase, we propose the MDCL model for MSA, which is shown in Fig. 1. Firstly, two single-modal encoders are used to obtain representations for text and images, integrating these single-modal representations through attention-based modal fusion. Next, these integrated representations serve as inputs for the Momentum Distillation model, facilitating modal alignment and image-text matching. By constructing contrastive losses using pseudo-targets generated by the distillation model, we then proceed to learn and minimize the distance between text and image representations extracted from single-modal encoders, thereby achieving semantic alignment between modalities. Finally, following a similar approach, negative samples are constructed for image-text matching training, further enhancing the learning of correlations between different modal data.

Previous studies have shown that linguistic modalities contribute more information to multimodality compared to images, so we use a more complex text model text *T* is encoded by the text encoder to obtain representation M_T, which is shown as Equation 1, and image *I* is encoded by the image encoder (Resnet152 used) to obtain representation M_I, which is shown as Equation 2.

$$M_T = BERT(t) \tag{1}$$

$$M_I = ResNet(i) \tag{2}$$

The models we used are respectively pre-trained on large-scale textual and visual datasets. Simple feature fusion is ineffective for sentiment analysis because the task of sentiment analysis requires a model with the ability to extract high-level semantic information about a graphic. Image feature M_I, is spliced with text feature M_T in sequence dimension to obtain M_{IT}. We utilize the BERT layer with attention mechanism for inter-modal fusion, as shown in Equation 3.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{3}$$

Where *Q, K, V* are obtained from the input features M_{IT} by different linear transformations, d_k is the dimension of the *K*, *M* is the attention mask. The attention mask *M* consists of two parts. One part is obtained by the text encoder BERT, which identifies the effective part of the text sequence, while the image attention mask is artificially set to an all '1' vector, indicating that the image is considered all parts, the attention mask *M* is then converted and scaled as Equation 4 to convert it into a form suitable for the attention mechanism.

$$M = (1.0 - M) * (-10000) \tag{4}$$

For the fused multimodal representation, we use two loss functions for modal alignment and modal matching.

3.2 Loss of alignment objective

Inspired by [20], we use momentum model to generate image, text features similar to the main model, and in this way implement several features mentioned below. The momentum model performs parameter updates by delayed replication of the parameters of the main model, rather than by gradient descent. This makes its feature representation smoother, more stable, and occupies less computational resources. The update of the momentum model follows the following Equation 5.

$$\theta_m = \alpha \cdot \theta_m + (1 - \alpha) \cdot \theta \tag{5}$$

where θ_m is a parameter of the momentum distillation model, θ is a parameter of the main model, and α is the momentum coefficient. As shown in the lower right part of Fig. 1, the momentum model maintains two feature queues storing historically updated momentum features T_m^{all} and I_m^{all}. Following [32] we define the hybrid target similarity matrix in the following form Equations 6 and 7.

$$S_{targets}^{i2t} = \alpha \times softmax(S(I, T_m^{all})) + (1 - \alpha) \times S_{ones} \tag{6}$$

$$S_{targets}^{t2i} = \alpha \times softmax\big(S(T, I_m^{all})\big) + (1 - \alpha) \times S_{ones} \tag{7}$$

The image-to-text contrastive loss is defined below Equation 8, and text-to-image contrastive loss is defined below Equation 9, the above two loss functions achieve the function of eliminating heterogeneity between text and image by narrowing the distance between text samples (image samples), and generated image pseudo positive samples (generated text pseudo positive sample). Also, the final image-text contrastive loss is Equation 10. The content of this section is in the lower right side of Fig. 1.

$$loss_{i2t} = -\frac{1}{N}\sum \log\left(\frac{\exp\big(S(I, T_m^{all})\big)}{\sum \exp\big(S(I, T_m^{all})\big)}\right) * S_{targets}^{i2t} \tag{8}$$

$$loss_{t2i} = -\frac{1}{N}\sum \log\left(\frac{\exp\big(S(T, I_m^{all})\big)}{\sum \exp\big(S(T, I_m^{all})\big)}\right) * S_{targets}^{t2i} \tag{9}$$

$$loss_{ita} = \frac{1}{2(loss_{i2t} + loss_{t2i})} \tag{10}$$

3.3 Loss of matching objective

The image-text matching loss (ITM-loss) learns whether the image and text are matched or not, it will minimize the distance between the text and its matching image while maximizing the distance between the text and its mismatched image. For the selection of negative samples, we first use the previous hybrid target similarity matrix, which is calculated according to the following Equations 11 and 12 to get its probability distribution matrix. By setting the diagonal elements of the probability distribution matrix to '0' to exclude the influence of positive pairs of samples, so as to select the negative samples corresponding to the text or image.

$$P(I_i|T_j) = \frac{\exp(S_{ij})}{\sum_{k=1}^{n} \exp(S_{kj})} \quad (11)$$

$$P(T_j|I_i) = \frac{\exp(S_{ij})}{\sum_{k=1}^{m} \exp(S_{ik})} \quad (12)$$

The image-text matching loss is computed by the cross-entropy loss function such as Equation 13, where y_i represents the ground truth label, $\hat{y}_i$ is the matching probability calculated by the model. For positive samples, we want this probability to be close to 1; for negative samples, we want this probability to be close to 0. By optimizing ITM-loss, the model learns how to fuse features from different modalities more efficiently, thus improving the understanding of the integrated information. The content of this section is on the upper right side of Fig. 1.

$$loss_{itm} = -\frac{1}{N}\sum_{i=1}^{N} [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (13)$$

Finally, the multimodal synergy loss ($loss_{MSL}$) is set as Equation 14, where λ_1, λ_2 are hyper-parameters for weights of the losses.

$$loss_{MSL} = \lambda_1 loss_{ita} + \lambda_2 loss_{itm} \quad (14)$$

4. Experiments

4.1 Dataset

In order to test the ability of our model, we tested on two widely used public dataset MVSA-single and MVSA-multi. The MVSA dataset was collected from the social media platform twitter. The MVSA dataset was collected from the social media platform twitter. The MVSA-single dataset includes 4869 image-text pairs. Each image and text pair are annotated by one annotator, and MVSA-multi contains 19, 600 image and text pairs. Each image and text pair are annotated by three annotators, and each annotator independently annotates the image and text. For the MVSA-Single dataset, we delete a tweet if the image is completely opposite to the sentiment node of the text. And if the labels of the image and text do not have a neutral mode, the sentiment nodes of the tweet will be regarded as another modal sentiment graph. For example, if the image is labeled neutral and the text is labeled positive, then the tweet will be considered positive. For MVSA-multi, in addition to the above rules, since each tweet is annotated by three annotators, we set the label of each tweet to a majority vote among the three labels, that is, when at least two of the three annotators The label marked by the annotator is the final vote for the tweet. After being processed by the above rules, we obtained the MVSA-single data set of 4511 image-text pairs and the MVSA-multi data set of 16779 image-text pairs. The label statistics of the processed MVSA data set are as shown in Table 1.

Table 1. Sentiment polarity distribution of MVSA dataset.

Dataset	Positive	Neutral	Negatives	All
MVSA-S	2683	470	1358	4511
MVSA-M	9328	6359	1092	16779

4.2 Implementation details

We use the AdamW optimizer with weight decay set to 1e-2, batch size set to 64, parameters used to update the momentum model set to 0.995, pseudo target queue size maintained by the momentum model set to 16384, dropout set to 0.4. We use Bert-base and resnet152 pre-training parameters to initialize the text and image encoders in the model, λ_1=0.1, λ_2=0.2. Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU for 50 epochs of training.

4.3 Compared Methods

To evaluate the performance of our model, we compare it with the following model:

CLIP [21]: A contrastive learning model proposed by OpenAI in 2021, which has achieved state-of-the-art performance on several NLP tasks. The results shown in Table 2 were obtained from [4]. **ResNet-50** [22]: A pre-trained and fine-tuned model on the image only task. **Multiengine** [23]: A semantic network used for MSA. The results shown in Table 2 were reimplemented in [24]. **OSDA**

Table 2. The metrics of accuracy and F1-score on two datasets.

Modality	Model	MVSA-S		MVSA-M	
		ACC	F1	ACC	F1
Image	CLIP	-	-	66.13	51.90
	ResNet-50	64.67	61.55	61.88	60.98
	OSDA	66.75	66.51	66.62	66.23
Text	CNN-T	68.19	55.90	65.64	57.66
	BiLSTM-T	70.12	65.06	67.90	67.90
	BERT	71.11	69.70	67.59	66.24
Text+Image	MultiSentiNet	63.27	59.12	63.08	59.12
	CNN-Multi	66.30	64.19	-	-
	FENet-BERT	69.02	67.30	68.61	65.80
	MultiSentiNet-M	69.84	69.63	68.86	68.11
	ALBEF	-	-	69.86	56.47
	Se-MLNN	72.09	70.03	65.42	59.22
	Co-Memory-M	70.51	70.01	69.92	69.83
	MVAN-M	72.98	72.98	**72.36**	72.30
	OURS	**73.56**	70.71	70.66	68.27

[25]: An image sentiment analysis model based on multiple views. **MultiSentiNet-M** [23]: A visual-feature-guided LSTM with extract words that were important to text sentiment and then aggregated the text representation, image object features and scene features. **FENet-BERT** [24]: A fusion extraction network model for MSA. The results in Table 2 were reimplemented in [26]. **Se-MLNN(CI)**[26]: A network combines several visual features with contextual text features to predict the overall sentiment accurately. **CNN-Multi** [27]: uses two independent CNN architectures to learn the features of text and images and utilizes them as the input of another CNN for MSA. **CNN-T** [28]: A text sentiment analysis model based on CNN. **BiLSTM-T** [29]: A text sentiment analysis model based on a BiLSTM. **Co-Memory-M** [30]: A model using co-memory network to iteratively model the interactions between visual content and text for MSA. **MVAN-M** [25]: A network use of memory networks that are constantly updated to obtain deep semantic information about text and images. **ALBEF** [31]: A multi-modal pre-trained model that achieved state-of-the-art results on several multimodal tasks.

4.4 Results and analysis

The experimental results of the baseline approach with our model are shown in the Table 2, the results show that our model has the best results or has shown informative performance on the MVSA-S and MVSA-M datasets, according to the comparison of the results of the unimodal model and the multimodal model it can be found that the multimodal model generally outperforms the unimodal model on the MSA task, and that the second modality implies affective information which is indeed beneficial for the second modality contains emotional information that is indeed beneficial for the model to learn more comprehensive emotional information. Comparing the results of the unimodal model between the text and picture modalities, the text model generally shows better results than the picture modality. Thus, it can be intuitively demonstrated that text modality plays a more important role than image modality for the sentiment analysis task, which is contrary to the experience of the multimodal pre-training task[32], and we believe that this may be related to the way people express their emotions on social media, where people more often use text to express themselves in response to images, and thus text aggregates more sentiment information. The models that underwent modality alignment were better than those that did not, confirming that modality alignment reduces semantic differences between modalities and mitigates heterogeneity between modalities. Experimental performance observations on pre-trained models that have not been adjusted for the sentiment analysis task (e. g., ALBEF) yield that the performance of such pre-trained models is comparable to that of unimodal models (e. g., BERT) because they capture only semantic multimodal information while ignoring sentiment signals.

4.5 Ablation studies

We conducted ablation studies on the MVSA-S and MVSA-M datasets to illustrate the role of ITA-loss versus ITM-loss in our model. We removed ITM-loss and ITA-loss separately as well as the performance after both were removed. The results of the ablation study are presented in Table 3. ALL means that the model has all modules including, -ITA-ITM means model without ITA-loss module and ITM-loss module, -ITA means model without ITA-loss module, -ITM means model without ITM-loss module. The ablation study shows that the full version of MDCL achieves the best performance, which is marked in bold in Table3.

We first removed ITA-loss and retained ITM-loss. The results indicate that the alignment of intermodal features facilitates the model's ability to integrate cross-modal features, confirming the critical role of ITA-loss in improving the model's deep feature understanding. Then we remove the ITM-loss and retain the ITA-loss. The results show that by learning for both positive and negative sample pairs, the model is able to accurately match and correlate the semantic content between images and texts. The more significant improvement in the results also suggests that the model's understanding of the high-level semantics of multimodal information will have a more important effect on the sentiment categorization task. Removing both the graph alignment module and the graph matching module adversely affects the model. This indicates that both modules are effective for MSA.

Table 3. Results of ablation studies.

Dataset	Model	ACC	F1
MVSA-S	**ALL**	**73. 56**	**70. 71**
	-ITA-ITM	68. 10	66. 93
	-ITA	71. 09	68. 25
	-ITM	70. 44	69. 27
MVSA-M	**ALL**	**70. 66**	**68. 27**
	-ITA-ITM	68. 10	66. 93
	-ITA	67. 16	66. 23
	-ITM	66. 98	66. 02

5. Conclusions

In this paper, we proposed a multimodal data processing model that integrates momentum distillation and contrastive learning. This model combined the strengths of both techniques, leading to a more comprehensive and effective enhancement of the performance of deep learning models. Furthermore, model testing and ablation experiments were conducted on public datasets of MVSA-single and MVSA-multi. The experimental results demonstrated the superiority of our model in multimodal sentiment analysis, highlighting the beneficial contributions of the introduced dual modules in improving the efficiency of multimodal sentiment analysis.

As the future works, we are to: (1) Fine-tuning hyperparameters to optimize the performance of the multimodal sentiment analysis model, ensuring better adaptability to different datasets and scenarios; (2) Further research might delve into enhancing the scalability and efficiency of the proposed model, making it applicable to larger datasets and real-time applications.

Acknowledgement

This work was supported by JSPS KAKENHI under Grant 20H04284 (Grant-in-Aid for Scientific Research (B)), National Natural Science Foundation of China under Grant 72174079, Jiangsu Philosophy and Social Science (24YYC002), Lianyungang Philosophy and Social Science (24LKT0041), General Project of Philosophy and Social Science Research in Universities of Jiangsu Province (2023SJYB1829), Project Haiyan-2023, the Research Start-up Fund of JOU under Grant KQ23032.

References

[1] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis”, Image Vis. Comput., vol. 65, no. 1, pp. 3-14 (2017)

[2] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, and T. Sun, “Sentiment Analysis: Comprehensive Reviews, Recent Advances, and Open Challenges”, IEEE Trans. Neural Networks and Learning Systems, pp. 1-21 (2023)

[3] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis”, Proc. EMNLP, pp. 1103–1114 (2017)

[4] J. Ye, J. Zhou, J. Tian, R. Wang, J. Zhou, T. Gui, Q. Zhang, and X. Huang, “Sentiment-aware multimodal pre-training for multimodal sentiment analysis”, Knowledge-Based Systems, vol. 258, 110021 (2022)

[5] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web”, Proc. ICMI, pp. 169–176 (2011)

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition”, Neural Computation, vol. 1, no. 4, pp. 541–551 (1989)

[7] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory”, Neural Computation, vol. 9, no. 8, pp. 1735-1780 (1997)

[8] X. Yang, S. Feng, D. Wang, and Y. Zhang, “Image-Text Multimodal Emotion Classification via Multi-View Attentional Network”, IEEE Trans. Multimedia, vol. 23, pp. 4014-4026 (2021)

[9] T. Jiang, J. Wang, Z. Liu, and Y. Ling, “Fusion-Extraction Network for Multimodal Sentiment Analysis”, Proc. PAKDD, pp. 785–797 (2020)

[10] D. Chen, W. Su, P. Wu, and B. Hua, “Joint multimodal sentiment analysis based on information relevance”, Inf. Process. Manage., vol. 60, no. 2, Art. no. 103193 (2023),

[11] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multimodal Transformer for Unaligned Multimodal Language Sequences”, Proc. Annual Meeting of the Association for Computational Linguistics, pp. 6558–6569 (2019)

[12] S. Mai, Y. Zeng, S. Zheng, and H. Hu, “Hybrid Contrastive Learning

of Tri-Modal Representation for Multimodal Sentiment Analysis", IEEE Trans. Affect. Comput., vol. 14, no. 3, pp. 2276–2289 (2023)

[13] Z. Lin, B. Liang, Y. Long, Y. Dang, M. Yang, M. Zhang, and R. Xu, "Modeling Intra- and Inter-Modal Relations: Hierarchical Graph Contrastive Learning for Multimodal Sentiment Analysis", Proc. COLING, pp. 7124–7135 (2022)

[14] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, "CrossCLR: Cross-Modal Contrastive Learning for Multi-Modal Video Representations", Proc. ICCV, pp. 1450–1459 (2021)

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network", Available: http://arxiv.org/abs/1503.02531 (2015)

[16] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born Again Neural Networks", Proc. ICML, vol. 80, pp. 1607–1616 (2018)

[17] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, "Online knowledge distillation via collaborative learning", Proc. CVPR, pp. 11020–11029 (2020)

[18] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language", Proc. ICML, pp. 1298–1312 (2022)

[19] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation", Proc. ICCV, pp. 1921–1930 (2019)

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning", Proc. CVPR, pp. 9729–9738 (2020)

[21] A. Bondielli, L. C. Passaro, and others, "Leveraging CLIP for Image Emotion Recognition", CEUR Workshop Proceedings, vol. 3015 (2021)

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", CVPR, pp. 770–778 (2016)

[23] N. Xu and W. Mao, "Multisentinet: A deep semantic network for multimodal sentiment analysis", Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2399–2402 (2017)

[24] T. Jiang, J. Wang, Z. Liu, and Y. Ling, "Fusion-extraction network for multimodal sentiment analysis", PAKDD, pp. 785–797 (2020)

[25] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network", IEEE Transactions on Multimedia, vol. 23, pp.4014–4026 (2020)

[26] G. S. Cheema, S. Hakimov, E. Müller-Budack, and R. Ewerth, "A fair and comprehensive comparison of multimodal tweet sentiment analysis methods", Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, pp. 37–45 (2021)

[27] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis", NLPCC, pp. 159–167 (2015)

[28] Y. Kim, "Convolutional neural networks for sentence classification ", EMNLP, pp. 1746–1751 (2014)

[29] P. Zhou et al., "Attention-based bidirectional long short-term memory networks for relation classification", Proc. 54th Annu. Meeting Assoc.Comput. Linguistics (volume 2: Short papers), pp. 207–212 (2016)

[30] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis", ACM SIGIR, pp. 929-932 (2018)

[31] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation", Advances in Neural Information Processing Systems, pp.9694–9705 (2021)

[32] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision", PMLR, pp. 5583–5594 (2021)

〈Author Biography〉

GAN Quan: Ph.D., Yamaguchi University; Lecturer, Jiangsu Ocean University; Multimodal fusion and system modeling.
WANG Chen: B.E., Anhui University of Science and Technology; Natural language processing and computer vision.
WU Jiaying: Ph.D., Waseda University; Lecturer, Jiangsu Ocean University; Computer vision and few-shot learning.
ZHONG Zhaoman: Ph.D., Shanghai University; Professor, Jiangsu Ocean University; Information retrieval and data analysis.
GE Qi-Wei: Ph.D., Hiroshima University; Professor, Yamaguchi University; Petri nets and combinatorics.
LIU Chuanxia: Ph.D., Yamaguchi University; Lecturer, Jiangsu Ocean University; Data analysis and cognitive psychology.