International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 12 Issue 3 March 2023 || PP 53-60

Multimodal Similarity Fusion Network for Multimodal Sentiment Analysis

Binghui Su[1], Zhenwen Sheng[2,*]
[1](College of Railway Transportation/ Hunan University of Technology, Zhuzhou)
[2](Shandong Xiehe University, Jinan)
*Correspondence: shengzhenwen@sdxiehe.edu.cn

ABSTRACT : *Multimodal sentiment analysis has been applied to many natural interaction scenarios, aiming to infer user emotional states from visual, audio, and linguistic modal data. Multiple models in multimodal sentiment analysis have been dedicated to exploring multimodal data fusion mechanisms for improving model performance. However, the multimodal data collected by different sensors are heterogeneous, which invokes tremendous challenges for multimodal data fusion and interactive operation. This paper proposes a new multimodal sentiment analysis model, MSFN, which aims to reduce heterogeneity differences among modalities and construct good intra-modal and inter-modal relationships. MSFN imposes two strategies of similarity learning on multimodal embedding representations. The first strategy is to add a generative adversarial loss function for learning the commonalities of different modalities and reducing modality differences. The second strategy is a distance metric function to reduce the distance between features with the same semantic embedding. Comprehensive experiments on public datasets show that our model outperforms the baseline models.*

KEYWORDS - *Data Fusion, Multimodal Sentiment Analysis, Generative Adversarial Network, Distance Metric*

Date of Submission: 14-03-2023 Date of Acceptance: 31-03-2023

I. INTRODUCTION

The research field of sentiment analysis in natural language processing (NLP) has accumulated a considerable amount of results. The Language-based unimodal sentiment analysis has made tremendous progress [1-3], whereas the pre-trained models with a large number of parameters such as BERT [4] and some of its variants [5, 6] continue to integrate the state-of-the-art (SOTA) technology pushed to new heights. Different from unimodal sentiment analysis, Multimodal Sentiment Analysis (MSA) utilizes relevant information extracted from multimodal data for comprehensive and comprehensive sentiment analysis [7].

Several models in MSA focus on exploring the complex fusion strategies between different modalities. The attention-based [8] and tensor fusion-based [9] models are examples. However, most of the above models do not consider that the data features of different modalities have inconsistent distributions and representations before fusion. This distribution difference between the modalities is termed the heterogeneity gap between the modalities, which seriously hinders the subsequent interactive operation of the multimodal data [7]. The mainstream model for bridging the heterogeneity gap is to map the features of different modalities into a common subspace for discussion. The models based on the relevance measures [10] and the ones that employ the generative adversarial networks for cross-modal translation [11] are examples. Although several models have been proposed for studying the common subspaces, there is still much room for improvement.

To construct good intra-modal and inter-modal relations, this paper proposes a multimodal similarity fusion network for multimodal sentiment analysis. The contributions of this paper can be summarized in the following two points.

1.1 A new MSA-oriented model, I2MCL, is proposed. It adopts two training strategies, multimodal adversarial loss function, and embedded feature distance metric loss function, for establishing good intra-modal and inter-modal relations.

1.2 Extensive comparative experiments over the popular benchmark dataset demonstrate that the MSFN model outperforms the baseline models with respect to the multiple evaluation metrics.

II.RELATED WORK

The relevant literature on MSA can be roughly categorized according to the methods used, viz., (1) the methods for learning complex fusion mechanisms and (2) the methods for learning common subspaces.

2.1 Methods for Learning Complex Fusion Mechanisms

Exploring the complex fusion mechanisms is an effective approach to addressing MSA, and several studies have focused on this approach. The work [9] has pioneered in obtaining the fusion tensors by employing multimodal features for outer product operations. Accordingly, the work [12] proposes Low-rank Multimodal Fusion (LMF) to reduce the computational cost of the outer product operations. The work [13] employs the outer product operation for fusion, but it divides the modality features into multiple local blocks before fusion to prevent the generation of high-dimensional fusion tensors. Furthermore, attention-based and gating-based fusion networks have been studied in several works. The work [14] proposes a Delta Memory Attention Network (DMAN) for cross-view interactions. The work [8] employs the components of the Multi-Attention Block (MAB) to discover the interaction between the modalities and improves LSTM so that the attention weights of different modalities are shared among the various LSTM components. The work [15] proposes Dynamic Invariant Specific Representation Fusion Network (DISRFN).

2.2 Methods for Learning Common Subspaces

Recently, an increasing number of scholars have studied the heterogeneity gap in MSA. The mainstream approach to solving this problem is to map the features of different modalities to a common subspace for discussion. Certain studies attempt to add cross-modal correlation and similarity constraints to the model. The work [10] learns the correlation between multiple modalities through Deep Canonical Correlation Analysis (DCCA). The works [16, 17] have reduced the distribution difference between modalities by minimizing the Maximum Mean Discrepancy (MMD) [18] and the Central Moment Discrepancy (CMD) [19] distance metrics. Furthermore, Generative Adversarial Networks (GANs) have been demonstrated for their powerful ability to transform data distributions and learn discriminative representations, and hence, successfully applied to aiding the encoders to learn common subspace representations [7]. The work [11] proposes an adversarial encoder-decoder-classifier for modality translation. The appeal methods only focus on the intra-modal or inter-modal, and the similarity relationship between the intra-modal and inter-modal is not well established. MSFN employs two training strategies to bridge modality differences while reducing the distance between features with the same semantic embedding within a modality.

III.METHODOLOGY

This section briefly defines the task settings of the MSA, followed by a description of the MSFN model. The task of MSA is to build a model to predict the emotional state of a video clip through a multimodal dataset. The input to the model is a set of multimodal datasets $\mathrm{X}=\{\mathrm{X}_a, \mathrm{X}_v, \mathrm{X}_l\}$ containing N segments, where audio, visual, and language modality data are denoted as $\mathrm{X}_a \in \mathbb{R}^{N\times d_a}$, $\mathrm{X}_l \in \mathbb{R}^{N\times d_l}$ and $\mathrm{X}_l \in \mathbb{R}^{N\times d_l}$, respectively, and $d_{m\in\{a,v,l\}}$ is the feature dimension of the corresponding modality. The primary task of the model is to extract and integrate task-related information from these multimodal data to form a unified embedding representation, then employ these representations to predict the continuous sentiment score label $y \in [-3,3]$.

3.1. Framework

According to Fig.1, the MSFN model framework has four main parts: Multimodal Data Embedding, Multimodal Similarity Learning, Multimodal Decoder, and Fusion Prediction Network.

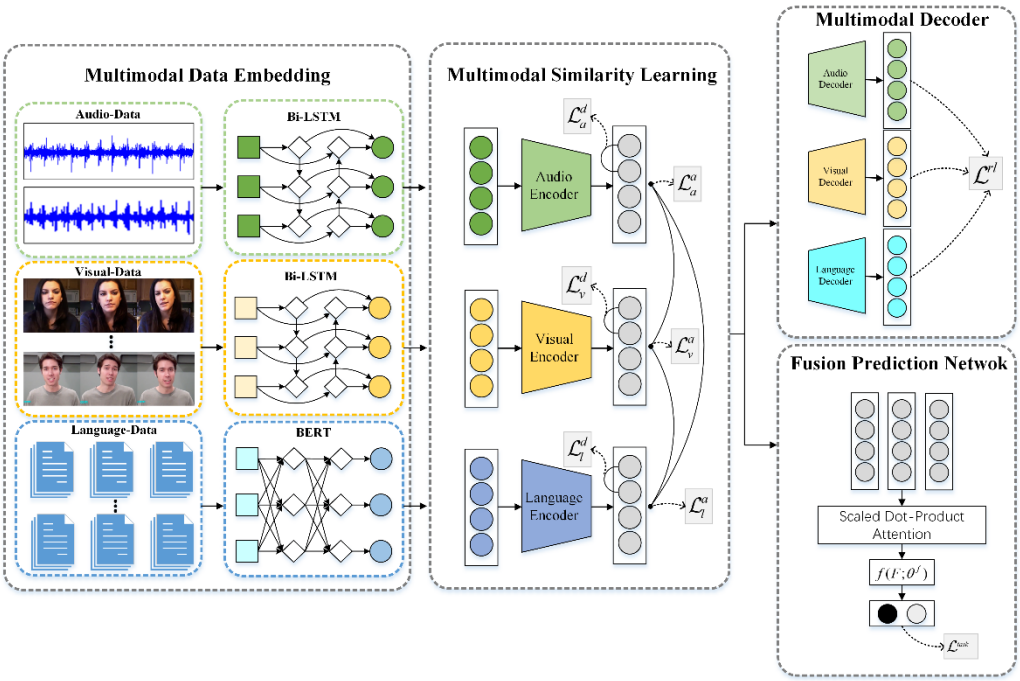

Figure 1. The overall structure of MSFN

3.2 Multimodal Data Embedding

For the audio and visual modalities, two LSTMs blocks have been utilized. The feature extraction was performed on the audio and visual modality data, and the output is the audio and visual modality embedding features, viz., $\mathbb{R}^{d_{m\in\{a,v\}}} \to \mathbb{R}^{d}$. The LSTMs block consists of two parts, viz., the bidirectional long short-term memory network Bi-LSTM and the fully connected Dense layer. Bi-LSTM was employed to extract the basic features of the audio modality and visual modality. The dense layer was used to receive data from Bi-LSTM and the embedded features of the LSTM end state, then unify the dimensions of the embedded features of different modalities as：

$$x_a = \mathrm{LSTMs}(\mathrm{X}_a;\theta_a^{lstm}) \tag{1}$$

$$x_v = \mathrm{LSTMs}(\mathrm{X}_v;\theta_v^{lstm}) \tag{2}$$

where θ_a^{lstm} and θ_v^{lstm} denote the parameters of LSTMs blocks in audio and visual modalities, respectively.

For the language modality, the BERT model is the latest technological achievement in natural language processing. It has achieved appreciable results in downstream tasks such as sentence pair classification, single sentence classification, question answering, and single sentence tagging tasks. A pre-trained transformer model BERT was employed for the feature extraction of language modality data. Furthermore, like LSTMs, a fully connected Dense layer was connected to the output layer of BERT to form the BERTs component: $\mathbb{R}^{N\times d_l} \to \mathbb{R}^{N\times d}$, the output is the language modality embedding feature x_l, as：

$$x_l = \mathrm{BERTs}(\mathrm{X}_l;\theta_l^{bert}) \tag{3}$$

where θ_l^{bert} is the parameter of the BERTs block.

3.3 Multimodal Similarity Learning

Although the multimodal embedded features $x_{m\in\{a,v,l\}}$ are uniform in dimension, their distribution in the feature space is inconsistent, i.e., there is a heterogeneity gap in the features of different modalities that hinders the subsequent fusion and interaction of different modalities. Inspired by generative adversarial networks [20] and similarity metric functions [18, 19], this paper combines a multimodal adversarial loss function and an embedded feature distance metric loss function. They are used to bridge the heterogeneity gap between modalities while increasing the similarity between the same semantic features. For each unimodal modality $m \in \{a,v,l\}$, we separately construct the encoder, which maps the multimodal embedding features x_m as input to a shared latent feature subspace as:

$$h_m = G_m\left(x_m;\theta^{G_m}\right), m \in \{a,v,l\} \tag{4}$$

where G_m parameterized by θ_{G_m} is a unimodal modality encoder, besides being a multimodal feature mapped to a shared subspace.

3.3.1 Multimodal Adversarial Loss

The multimodal adversarial loss aims to learn common representations between different modal data and reduce the heterogeneity gap between modalities. The generator G_m encodes multimodal features x_m in an attempt to trick the discriminator D_m into mapping the multimodal features into a common subspace.

For each modality, the corresponding modality generator network G_m and discriminator network D_m are set up. The discriminator network judges the corresponding modal input as true and other modal inputs as false. For example, for the language modality, the discriminator network D_l discriminates the language modality feature xl as true, while the audio modality feature D_a and visual modality feature D_v are discriminated as false. The multi-modal adversarial loss functions are as follows:：

$$\mathcal{L}_l^a = \mathbb{E}_{x_m \sim P_{x_m}}[\log(D_l(G_l(x_l))) - \log(D_l(G_v(x_v))) - \log(D_l(G_a(x_a)))] \tag{5}$$

$$\mathcal{L}_a^a = \mathbb{E}_{x_m \sim P_{x_m}}[\log(D_a(G_a(x_a))) - \log(D_a(G_v(x_v))) - \log(D_a(G_l(x_l)))] \tag{6}$$

$$\mathcal{L}_v^a = \mathbb{E}_{x_m \sim P_{x_m}}[\log(D_v(G_v(x_v))) - \log(D_v(G_a(x_a))) - \log(D_v(G_l(x_l)))] \tag{7}$$

$$\mathcal{L}^a = \frac{1}{3}\left(\mathcal{L}_l^a + \mathcal{L}_a^a + \mathcal{L}_v^a\right) \tag{8}$$

Through the above loss function, the generator and the discriminator have trained adversarially through minimax games to map multimodal features into a common subspace.

3.3.2 Distance Metric Loss

Distance metric loss functions have been proposed for computing distributed sample problems independently [18, 19]. In this paper, the triplet loss function [21] is used to optimize the intra-modal embedding feature space, so that features with the same semantics are closer to each other, thereby maintaining the similarity of intra-modal features. Compared with the previous distance measurement loss function, the triplet loss function has more definitions of anchor points, positive samples and negative samples. Positive samples are the same as the anchor class, and negative samples are different from the anchor class. In the embedding feature space, the distance from the positive sample to the anchor point should be shorter than the negative sample by a boundary value δ . For each modality, the following loss function is performed:

$$\mathcal{L}_m^d = \sum_{\substack{a,p,n \\ y_a = y_p \neq y_n}} (\delta + d_{a,p}^m - d_{a,n}^m) \tag{9}$$

$$\mathcal{L}^d = \frac{1}{3}\left(\mathcal{L}_l^d + \mathcal{L}_a^d + \mathcal{L}_v^d\right) \tag{10}$$

where $d_{a,p}^m$ and $d_{a,n}^m$ are distance metric functions used to calculate the Euclidean distance between positive and negative sample pairs.

3.4 Multimodal Decoder

Multimodal features have the risk of information loss during the feature mapping process [17, 22]. Therefore, a decoder was designed for unimodal modalities to reconstruct the multimodal input to reduce the impact of this risk. Particularly, for each unimodal modality $m \in \{a,v,l\}$, we construct the decoder for the reconstruction of the feature h_m , and the reconstruction loss function is given as：

$$\hat{x}_m = E_m\left(h_m;\theta^{E_m}\right) \tag{11}$$

$$\mathcal{L}^{rl} = \frac{1}{3}\sum_{m\in\{a,v,l\}} \left\| x_m - \hat{x}_m \right\|_2^2 \tag{12}$$

3.5 Fusion Prediction Network

The learned multimodal features are first fed into a self-attention layer. The Scaled Dot-Product Attention [23] was introduced to calculate the unimodal modality：

$$F_m = Softmax(\frac{Q_m K_m^{\ T}}{\sqrt{d_k}})\mathrm{V}_m, m \in \{a,v,l\} \tag{13}$$

where Q_m, K_m, and V_m are obtained from the unimodal modality through different linear transformations, viz.：

$$Q_m = h_m w_m^q, K_m = h_m w_m^k, \mathrm{V}_m = h_m w_m^v, m \in \{a,v,l\} \tag{14}$$

where $w_m^q \in \mathbb{R}^{d\times d_q}$, $w_m^k \in \mathbb{R}^{d\times d_k}$, and $w_m^v \in \mathbb{R}^{d\times d_v}$ represent the linear transformation matrices of query, key, and value, respectively, and $d_q = d_k = d_v = d$.

Finally, the F_m is connected to obtain the multimodal fusion feature $F = [F_a, F_v, F_l]$, which is sent to a fusion prediction network $\hat{y} = f(F;\theta^f)$ composed of fully connected layers, and the task loss is defined as:

$$\mathcal{L}^{task} = \sum_{i=1}^{M} \left\| y_i - \hat{y}_i \right\|_2^2 \tag{15}$$

3.6 Optimization

The total loss of the model is weighted by task, multimodal adversarial, distance metric and reconstruction loss as follows:

$$\mathcal{L} = \mathcal{L}^{task} + \alpha\mathcal{L}^a + \beta\mathcal{L}^d + \gamma\mathcal{L}^{rl} \tag{16}$$

where α, β, and γ are the respective loss weight hyperparameters.

IV.EXPERIMENT

4.1 Dataset

The CMU-MOSI [] and CMU-MOSEI [] datasets collect video collections from online sharing websites, containing 296 and 3228 video clips, respectively, and each clip contains three modalities: language, audio, and vision. Each segment has a sentiment label in the range [-3,3]. The dataset can be found at: https://github.com/A2Zadeh/CMU-MultimodalSDKF.

4.2 Model Configuration

The MSFN model is built on the Pytorch platform, using the grid search method to adjust the hyper-parameters in the model, thus saving the appropriate hyper-parameters. Table 1 shows the final hyper-parameters of the two datasets determined with the grid search method, and Fig. 2 shows the component structure of MSFN. (1) d represents the unified dimension of the different modalities, and (2) α, β, and γ come from the formula (19). Further, (3) r, Batch_ Size, and Drop represent the Learning Rate, batch size, and dropout rate in the iterative optimization, respectively. (4) LSTM represents the hidden layer dimension of the Bi-LSTM end state, and (5) BERT represents the output dimension of the BERT model. (6) Layer-Norm represents the dimension of the batch normalization layer.

Table 1： Hyper-parameters in the MSFN model

Hyper-param	MOSI	MOSEI
d	128	256
α	0.2	0.4
β	0.5	0.3
γ	0.2	0.1
r	1e-4	1e-4
Batch_Size	64	16

Drop	0.1	0.1

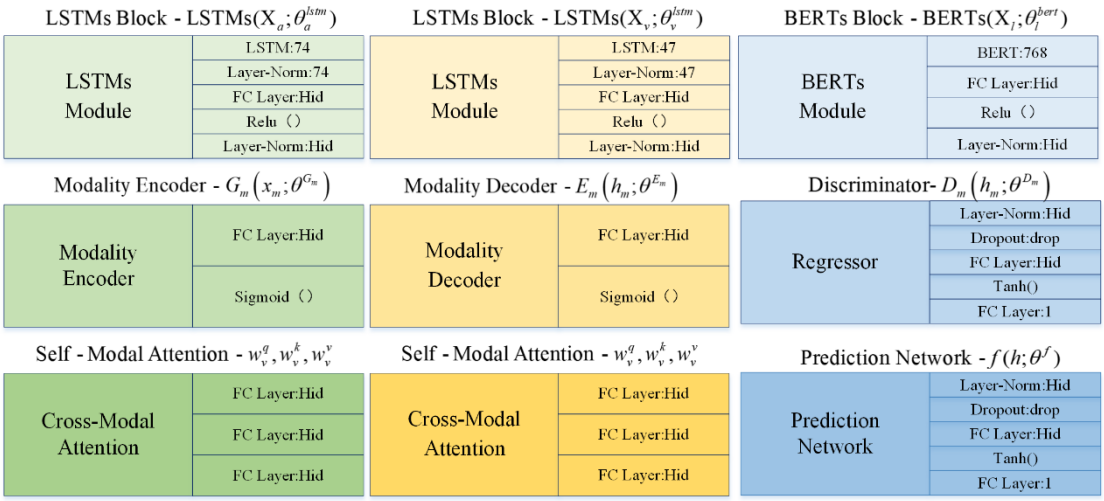

Figure 2. Component parameter settings of the MSFN model.

4.3 Benchmark Model and Comparison Results

Table 2：Comparison Experiments of Multimodal Models in MOSI

Model	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑
MFN	77.3%	77.4%	1.042	0.615	30.6%
LMF	77.3%	77.3%	1.005	0.622	35.4%
TFN	78.2%	78.3%	0.987	0.615	34.5%
TFN(B) ▽	80.8%	80.8%	0.901	0.698	34.9%
LMF(B) ▽	82.5%	82.5%	0.917	0.695	33.2%
ICCN(B) ▽	83.1%	83.0%	0.862	0.714	39.0%
ARGF(B)	83.4%	83.6%	0.826	0.740	37.9%
MISA(B)	83.1%	83.0%	**0.796**	0.749	**43.4%**
MSFN	**83.4%**	**83.5%**	0.802	**0.756**	43.1%

Note: (B) indicates that the BERT model is used for feature extraction and encoding of language modality; ▽ come from [10].

Table 3： Comparison Experiments of Multimodal Models in MOSEI

Model	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑
TFN	76.2%	76.1%	0.722	0.510	43.2%
MFN	76.6%	76.7%	0.727	0.511	43.9%
LMF	76.2%	76.4%	0.712	0.540	45.1%
TFN(B)$^{\triangledown}$	82.6%	82.1%	0.593	0.700	50.2%
LMF(B)$^{\triangledown}$	82.0%	82.2%	0.623	0.677	48.0%
ICCN(B)$^{\triangledown}$	84.2%	84.4%	0.565	0.713	51.6%
MISA(B)	84.1%	83.6%	0.563	**0.755**	50.8%
ARGF(B)	84.2%	**84.7%**	0.572	0.740	51.2%
MSFN	**84.5%**	84.4%	**0.560**	0.747	**51.6%**

Note: (B) indicates that the BERT model is used for feature extraction and encoding of language modality; $\triangledown$ come from [10].

Table 2 and Table 3 show the comparative experimental results of MSFN and other baseline models on the two datasets. MSFN achieves the best performance on most regression and classification metrics, surpassing previous strong baseline models. Specifically, MSFN achieves 0.3% binary accuracy on the MOSI and MOSEI datasets compared to the previous baseline model. Certain studies [4, 17, 19] have shown that BERT is superior when used as a language modality feature extractor compared to traditional methods. For a fair comparison, Tables 2 and 3 show some models that also use BERT as a language modality feature extractor. MSFN outperforms models using complex fusion mechanisms such as TFN, LMF, etc. Therefore, adding a multimodal adversarial loss is important because the heterogeneity gap hinders the fusion interaction of different modalities. Furthermore, the MSFN model also outperforms models that also learn common subspaces (such as ARGF), suggesting that adding a distance metric loss helps preserve the similarity of features within a modality. In the subsequent ablation experiments, this paper further illustrates the impact of multimodal contrastive loss and distance metric loss functions on the model.

4.4 Ablation Experiment

We have designed the ablation experiments on MOSI datasets, including the modality ablation experiments and the multimodal representation learning and ablation.

Table 4: Ablation experiments

Model	Acc-2↑	F1↑	MAE↓	Corr↑	Acc-7↑
1 MSFN	**83.4%**	**83.5%**	**0.802**	0.756	43.1%
2 (-) $\mathcal{L}^a$	82.5%	82.6%	0.835	0.746	42.1%
3 (-) $\mathcal{L}^d$	82.2%	82.5%	0.821	0.743	41.2%
4 (-) $\mathcal{L}^{rl}$	83.1%	83.2%	0.833	**0.762**	43.%

Note: (-) indicates a variant model from which this factor has been removed.

Models 2-4 in Table 5 are loss function ablation models, representing MSFN variants with multimodal adversarial, distance metric, and reconstruction loss removed, respectively. According to the table above, the best experimental results are achieved when the model includes all losses. This shows that all added loss functions are profitable. Furthermore, the model is more sensitive to both $\mathcal{L}^a$ and $\mathcal{L}^d$. Specifically, removing $\mathcal{L}^a$ or $\mathcal{L}^d$ leads to a drop in binary classification accuracy of 0.9% and 1.2%, respectively. This shows that the designed loss function helps to build a similarity structure within and between modalities, thereby significantly improving the model performance. In addition, the model is not sensitive to reconstruction loss $\mathcal{L}^{rl}$, the possible reason is that other loss functions designed also help reduce the risk of information loss.

V. CONCLUSION

This work proposes a new model for multimodal sentiment analysis - MSFN. MSFN aims to reduce the heterogeneity difference between modalities and maintain the similar structure of features within a modality. Comprehensive experiments on two popular datasets show that MSFN outperforms strong baseline models. In addition, other experimental results in this design also show that well-designed multiple loss functions all contribute to the improvement of model performance. This paper only evaluates the performance of the MSFN model on public and widely used datasets. In our future work, we will use more realistic multimodal sentiment

analysis data or multimodal data from other domains to verify the generalization ability of our method and further improve the robustness of our method.

REFERENCES

[1]. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. In Emotional chatting machine: Emotional conversation generation with internal and external memory, Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[2]. Wu, H.; Gu, Y.; Sun, S.; Gu, X. In Aspect-based opinion summarization with convolutional neural networks, 2016 International Joint Conference on Neural Networks (IJCNN), IEEE: 2016; pp 3157-3163.

[3]. Wang, K.; Wan, X., Automatic generation of sentimental texts via mixture adversarial networks. Artificial Intelligence **2019,** 275, 540-558.

[4]. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 **2018**.

[5]. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V., Roberta: A robustly optimized bert pretraining approach. arXiv 2019. arXiv preprint arXiv:1907.11692 **2019**.

[6]. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; Le, Q. V., Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **2019,** 32.

[7]. Guo, W.; Wang, J.; Wang, S., Deep multimodal representation learning: A survey. IEEE Access **2019,** 7, 63373-63394.

[8]. Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.-P. In Multi-attention recurrent network for human communication comprehension, Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[9]. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.-P., Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250 **2017**.

[10]. Sun, Z.; Sarma, P.; Sethares, W.; Liang, Y. In Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, Proceedings of the AAAI Conference on Artificial Intelligence, 2020; pp 8992-8999.

[11]. Mai, S.; Hu, H.; Xing, S. In Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, Proceedings of the AAAI Conference on Artificial Intelligence, 2020; pp 164-172.

[12]. Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; Morency, L.-P., Efficient low-rank multimodal fusion with modality-specific factors. arXiv preprint arXiv:1806.00064 **2018**.

[13]. Mai, S.; Hu, H.; Xing, S. In Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing, Proceedings of the 57th annual meeting of the association for computational linguistics, 2019; pp 481-492.

[14]. Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. In Memory fusion network for multi-view sequential learning, Proceedings of the AAAI conference on artificial intelligence, 2018.

[15]. He, J.; Yanga, H.; Zhang, C.; Chen, H.; Xua, Y., Dynamic invariant-specific representation fusion network for multimodal sentiment analysis. Computational Intelligence and Neuroscience **2022,** 2022.

[16]. Huang, X.; Peng, Y.; Yuan, M., MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval. IEEE transactions on cybernetics **2018,** 50 (3), 1047-1059.

[17]. Hazarika, D.; Zimmermann, R.; Poria, S. In Misa: Modality-invariant and-specific representations for multimodal sentiment analysis, Proceedings of the 28th ACM international conference on multimedia, 2020; pp 1122-1131.

[18]. Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; Smola, A., A kernel two-sample test. The Journal of Machine Learning Research **2012,** 13 (1), 723-773.

[19]. Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; Saminger-Platz, S., Central moment discrepancy (cmd) for domain-invariant representation learning. arXiv preprint arXiv:1702.08811 **2017**.

[20]. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y., Generative adversarial networks. Communications of the ACM **2020,** 63 (11), 139-144.

[21]. Schroff, F.; Kalenichenko, D.; Philbin, J. In Facenet: A unified embedding for face recognition and clustering, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015; pp 815-823.

[22]. Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; Erhan, D., Domain separation networks. Advances in neural information processing systems **2016,** 29.

[23]. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I., Attention is all you need. Advances in neural information processing systems **2017,** 30.

Binghui Su. "Multimodal Similarity Fusion Network for Multimodal Sentiment Analysis." *International Journal of Engineering Science Invention (IJESI)*, Vol. 12(3), 2023, PP 53-60. Journal DOI- 10.35629/6734