ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413608

MULTIMODAL EMOTION RECOGNITION WITH CAPSULE GRAPH CONVOLUTIONAL BASED REPRESENTATION FUSION

Jiaxing Liu [1,3], *Sen Chen*[1], *Longbiao Wang*[1,*], *Zhilei Liu*[1,*], *Yahui Fu*[1], *Lili Guo*[1], *Jianwu Dang*[1,2,3]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]Pengcheng Laboratory, Shenzhen, China
{jiaxingliu,chensen, longbiao_wang, zhileiliu, fuyahui, liliguo}@tju.edu.cn, jdang@jaist.ac.jp

ABSTRACT

Due to the more robust characteristics compared to unimodal, audio-video multimodal emotion recognition (MER) has attracted a lot of attention. The efficiency of representation fusion algorithm often determines the performance of MER. Although there are many fusion algorithms, information redundancy and information complementarity are usually ignored. In this paper, we propose a novel representation fusion method, Capsule Graph Convolutional Network (CapsGCN). Firstly, after unimodal representation learning, the extracted audio and video representations are distilled by capsule network and encapsulated into multimodal capsules respectively. Multimodal capsules can effectively reduce data redundancy by the dynamic routing algorithm. Secondly, the multimodal capsules with their inter-relations and intra-relations are treated as a graph structure. The graph structure is learned by Graph Convolutional Network (GCN) to get hidden representation which is a good supplement for information complementarity. Finally, the multimodal capsules and hidden relational representation learned by CapsGCN are fed to multihead self-attention to balance the contributions of source representation and relational representation. To verify the performance, visualization of representation, the results of commonly used fusion methods, and ablation studies of the proposed CapsGCN are provided. Our proposed fusion method achieves 80.83% accuracy and 80.23% F1 score on eNTERFACE05'.

***Index Terms*—** multimodal emotion recognition, capsule networks, graph convolutional, VGG-16

1. INTRODUCTION

Emotional expression plays a vital role in interpersonal communication [1], and successfully detecting the emotional states has practical importance for artificial intelligence (AI). Emotion recognition has considerable prospects in sociable robotics, medical treatment, education quality evaluation, and many other human-computer interaction systems [2]. Especially current COVID situation, the mentioned products are more meaningful.

Humans express emotions in various ways, such as speech [3, 4], body gestures [5], facial expressions [6], and text [7]. The unimodal signal cannot fully convey the true intention. Different modality describes different aspects of the same emotion. Therefore, multimodal signals are more robust and more in line with human expression habits. In this work, we research on audio and video modalities which are the most common and effective ways for humans.

The key to the success of multimodal emotion recognition is the fusion of multimodal information. Information fusion methods are mainly divided into two categories [8]. One is early fusion (feature-level fusion). The source signals, or the extracted representations are concatenated as the fusion representations at the early stage. Tripathi et al. [9] used Bidirectional Long Short Term Memory (BLSTM) to extract the features and directly merged the extracted representations. Although early fusion methods have low computational complexity, the existence of redundancy reduces the effectiveness of information. The other type of fusion is late fusion (score-level fusion) [10, 11]. The extracted representations or unimodal results are fused at the late stage. Zhang et al. [12] introduced Deep Belief Networks (DBN) to fuse the audio and video representations before the classifier. Atmaja et al. [13] used support vector regression (SVR) to combine early and late fusion results which was one kind of multi-step score-level fusion. The disadvantage of late fusion is lacking representation complementarity between two modalities. One hybrid fusion method can be called model-level fusion which benefits from the powerful model to improve the performance. Huang et al. [14] introduced Transformer to fuse the audio and video representations. However, current fusion methods usually ignored the redundancy and complementarity of information between different modalities.

Capsule network (CapsNet) was proposed by Sabour et al. [15] which was quickly introduced to various research fields [16, 17, 18]. Under the premise of ensuring that information was not lost, CapsNet used a routing algorithm to distill the information into 'capsule'. Graph Convolutional Network (GCN) [19] was introduced to model the relational data and achieved good results in the field of Natural Language Processing (NLP) [20, 21, 22]. To reduce redundancy and enhance complementarity, we propose a novel fusion method, Capsule Graph Convolutional Network (CapsGCN) as shown in Fig. 2. Firstly, the CapsNet is introduced to encapsulate the audio and video representation into capsules respectively. The redundancy of the information in capsules is reduced by dynamic routing. Secondly, the extracted audio-video capsules with their inter and intra relations compose a relational graph. Then, GCN is introduced to learn hidden representations between audio-video capsules. The hidden representations are a good supplement for inforamtion complementarity. Finally, the audio-video capsules, and learned hidden representations are feed to the multihead self-attention [23].

The major contributions of this paper are summarized as: 1) A CapsGCN that considering multimodal information redundancy and information complementarity is proposed. 2) Attention mechanism is introduced to balance the contributions of different representations

[*] CORRESPONDING AUTHOR

and futher reduce redundancy.

2. MULTIMODAL EMOTION REPRESENTATION FUSION WITH CAPSGCN

2.1. Multimodal emotion recognition system

The proposed system as shown in Fig. 1 mainly consists of two parts, which are unimodal representation learning and multimodal representation fusion. The representation R_a and R_v are learned by two streams of the first part of the system. The multimodal representation fusion part is the proposed CapsGCN integrating attention, and the details are shown in Fig. 2. The learned fusion representation $\hat{R}_f$ is followed by a Flatten layer and a fully connected layer.

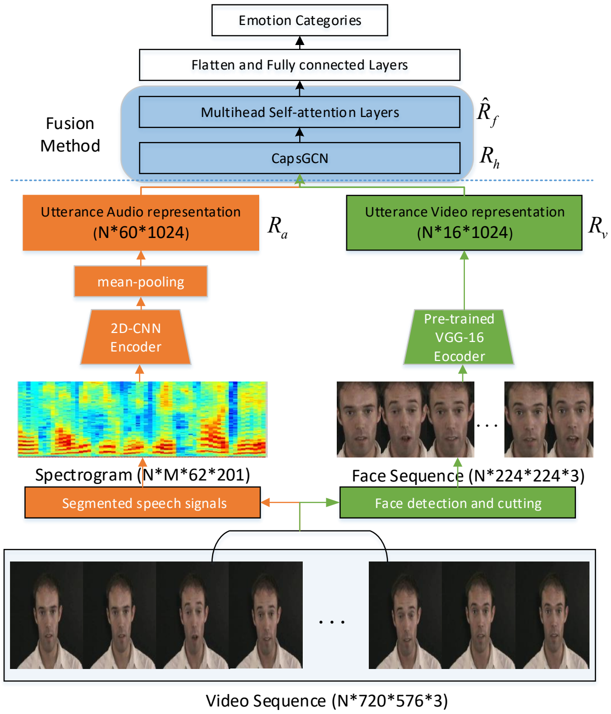

Emotion Categories
Flatten and Fully connected Layers
Fusion Method
Multihead Self-attention Layers
$\hat{R}_f$
CapsGCN
R_h
Utterance Audio representation (N*60*1024)
R_a
Utterance Video representation (N*16*1024)
R_v
mean-pooling
2D-CNN Encoder
Pre-trained VGG-16 Eocoder
Spectrogram (N*M*62*201)
Face Sequence (N*224*224*3)
Segmented speech signals
Face detection and cutting
Video Sequence (N*720*576*3)

Fig. 1. Multimodal emotion recognition system.

2.2. Unimodal representation learning

In this part, we use two streams to learn the unimodal representation. We divide the video sequence into audio data and image data, and the frame number is N. For audio data, the speech signals are divided into M segments with a 40ms overlap. Then the speech signals are transformed into spectrograms. We introduce common 2D-CNN to learn the audio representation and finally get the audio representation R_a $(N*60*1024)$. For image sequence, we use the OpenCV toolkit to detect the face and crop the face images whose samples are shown in Fig. 1. Due to data lacking consideration, at the video representation step, we introduce a pre-trained model VGG-16 [24] which is trained by ImageNet. In each fine-tuning epoch, 16 face images are randomly selected as the fine-tuning data. This training strategy is to prevent the redundancy of adjacent face images. Finally, we get the video representation R_v $(N*16*1024)$. The effectiveness of the learned representation R_a and R_v is verified in the experiments section.

2.3. CapsGCN based Representation Fusion

CapsNet outputs a vector instead of a single scalar value, which makes it be able to learn more obvious and complicated information. Dynamic routing method not only reduces information redundancy, but also avoids losing useful emotion information.

The c_{ij} are coupling coefficients which are determined by a “routing softmax” as shown in Eq. (1).

$$c_{ij} = \frac{\exp\left(d_{ij}\right)}{\sum_k \exp\left(d_{ik}\right)} \tag{1}$$

where the logits d_{ij} are the log prior probabilities that capsule i is coupled to capsule j

$$\hat{u}_{j|i} = W_{ij} u_i \tag{2}$$

$$s_j = \sum_i c_{ij} \hat{u}_{j|i} \tag{3}$$

In Eq.(4), v_j is the vector output of capsule j in layer l and s_j is its total input.

$$v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||} \tag{4}$$

We use the CapsNet to model the representation R_a and R_v, and get audio digitcap D_a and D_v:

$$\begin{aligned} D_a &= [v_0^a, v_1^a, ..., v_{k-1}^a, v_k^a] \\ D_v &= [v_0^v, v_1^v, ..., v_{k-1}^v, v_k^v] \end{aligned} \tag{5}$$

In Eq. (5), under the consideration of six emotion categories, we set $k = 6$ to capture simple and obvious information. The input data of the GCN is multimodal capsules R_{av} :

$$R_{av} = [v_0^a, v_1^a, ..., v_{k-1}^a, v_k^a, v_0^v, v_1^v, ..., v_{k-1}^v, v_k^v] \qquad v_i \in \nu \tag{6}$$

In Eq.(6), v_i represents the node, and the directed graph $G = (\nu, \varepsilon)$ and the edge is $(v_i, r, v_j) \in \varepsilon$. The hidden state in t-th layer:

$$h_i^{(t)} = \sigma\left(\frac{G_i^T (h_i^{(t-1)} W + b)}{\sum G_i}\right) \tag{7}$$

$\sigma()$ is an element-wise activation function, we use $ReLU()$ in this paper. $W$is the weight matrix and b is the bias. Finally we get the hidden representation R_h.

2.4. Attention Mechanism and Emotion Classification

In multimodal representation, the original independence of audio and video modal is also important. After the proposed CapsGCN, we concatenate R_{av} and R_h as fusion representation R_f. The fusion representation R_f is followed by multihead self-attention.

$$Att\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{D_K}}\right) V \tag{8}$$

In Eq. (8), the input matrix consists of Q, K, V which represents queries, keys, values respectively and the dimension of keys is D_K. Instead of performing a single calculation of Q, K, V, it is beneficial to linearly project the queries, keys, and values, h times with different learned linear projections. The h results are concatenated and

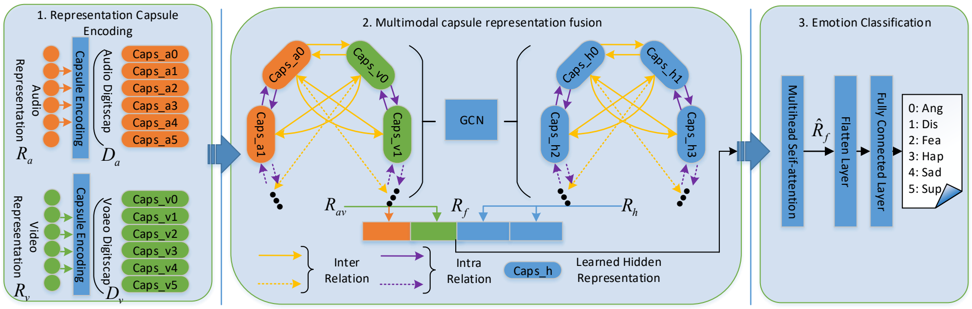


Fig. 2. The proposed model Capsule Graph Convolutional Network (CapsGCN) with Multihead Self-attention

once projected, resulting in the final output $M_h\left(Q, K, V\right)$ as shown in Eq. (9).

$$M_h\left(Q, K, V\right) = W\left(Att_1 + ... + Att_h\right) \tag{9}$$

in Eq. (9) $Q == K == V == R_f$. The R_f is calculated h times without sharing parameters and the h results are projected to $\hat{R}_f$ as shown in Eq. (10).

$$\hat{R}_f = M_h\left(R_f\right) \tag{10}$$

A balance on the source representation and hidden representation can be found by this attention mechanism to prevent individual emotions from being too prominent and further reduce redundancy.

3. EXPERIMENTS

3.1. Experiments Setup

3.1.1. *Datasets*

All experiments in this paper are conducted on eENTERFACE05' [25], which is an audio-visual database in English. This database contains six archetypal emotion categories, i.e., Anger, Disgust, Fear, Happiness, Sadness, and Surprise which is recorded by 42 subjects, coming from 14 different nationalities. Each of the subjects was told to listen to six successive short emotion-related stories. Only those subjects whose verbal and video reactions to each of the situations, as judged by two experts that the emotion was expressed in an unambiguous way, are included in the database. The audio data is recorded at 48kHz. The video data are processed using a 720x576 AVI format, and 25 frames per second.

3.1.2. *Audio and Video data Preprocessing*

The audio data is divided into equal-length segments with an overlap of 40ms, and each segment is converted into a 62 × 201 spectrogram. Because multiple segments can be generated from each audio sample, which will enlarge the training data. We use the OpenCV toolkit to do face detection and cutting jobs and obtain 224 × 224 face sequences.

3.1.3. *Audio and Video representation learning*

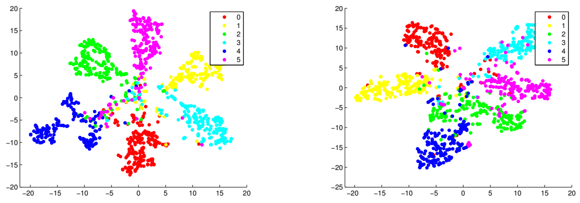
We use a commonly used 2D-CNN [26] as an audio representation learning method, and use the following mean-pooling layer to get the representation of each utterance. For video representation learning, we introduce pre-trained VGG-16 which is trained by ImageNet. We also set up two layers to follow the VGG-16 model. The video data fine-tunes the VGG model, and extracts the video representations. All experiments are conducted on speaker independent scheme.

3.2. Experimental Results and Analysis

To verify the effectiveness of the extracted audio and video representation and the proposed fusion method, we set up two groups of experiments to confirm the validity. The purposes of setting up the first group of experiments are that, one is to verify the effectiveness of preprocessing and representation learning, the other is to show the characteristics of unimodal information. The second group is to show the details of the proposed fusion method.

3.2.1. *Effectiveness of Audio and Video representation*

To observe the extracted audio and video representations, t-distributed stochastic neighbor embedding (t-SNE) [27] is introduced to visualize the six emotional categories as shown in Fig. 3.



(a) Audio representation (b) Video representation

Fig. 3. The t-SNE visualizations of unimodal representations.

We can find the intra relations in these two distributions are different. For example, in audio distribution there are a lot of points confusing together in the center, and in video distribution, Fear (green points) is very close to Suprise (purple points).

To verify the quantitative classification performance, we introduce DBN [12] and BLSTM [9] as the state-of-the-art comparative experiments. In these two comparative experiments, the DNB is introduced in score-level fusion manner, and BLSTM is in feature-level fusion manner. The experimental results of unimodal and multimodal experiments are shown in Table 1:

The visualizations and experimental results in Fig. 3 and Table 1 verify the effectiveness of the representation we extracted by

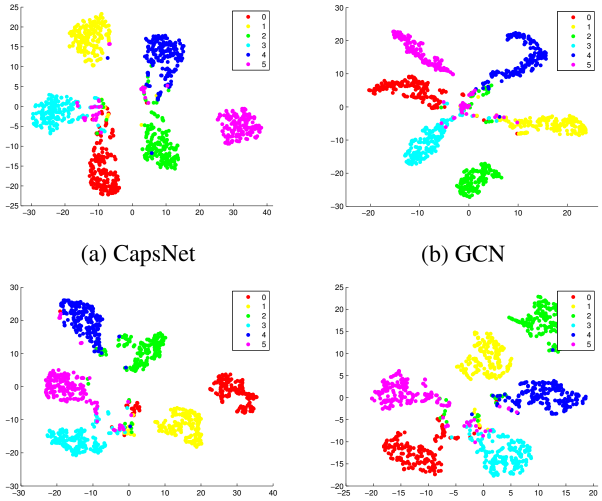
Table 1. Unimodal and multimodal experimental accuracy.

Representation	DBN [12](%)	BLSTM [9](%)
Audio	63.75	66.67
Video	57.08	58.33
Audio-Video	**69.17**	**71.67**

2D-CNN and pre-trained VGG-16. Furthermore, the fusion of audio and video representation could bring improvements whether it is a score-level fusion or a feature-level fusion. Although representations have been proven effective, these two state-of-the-art fusion methods do not bring good improvements. The performance of comparison algorithms Comparison algorithms are restrited by ignoring the importance of information redundancy and information complementarity.

3.2.2. *Validation of the proposed fusion method*

The following four t-SNE visualizations in Fig. 4 are the audio-video representation after (a)CapsNet, (b)GCN, (c)CapsGCN, and (d)CapsGCN integrating attention.



(a) CapsNet (b) GCN

(c) CapsGCN (d) CapsGCN_att

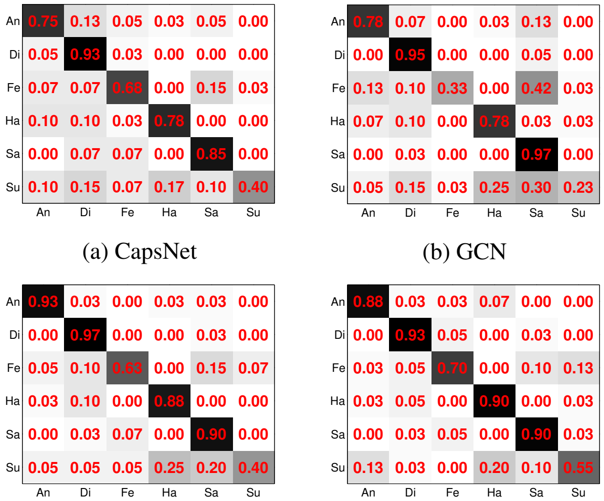
Fig. 4. The t-SNE visualizations of four fusion methods.

The distributions in (c)CaspGCN and (d)CapsGCN_att are clearer. To quantitatively analyze the proposed fusion methods, we provide the results of the ablation studies in Table 2, and the evaluation criteria are F1-score and accuracy. At the same time, four confusion matrices of the ablation studies are shown in Fig. 5.

Table 2. Ablation studies of proposed model.

Accuracy(%)	F1(%)	CapsNet	GCN	Attention
72.92	71.94	✓		
67.08	63.67		✓	
78.33	77.04	✓	✓	
80.83	**80.23**	✓	✓	✓

Observing Table 2 and Fig. 5, three phenomena can be found. The first one is that CapsNet performs better than GCN. This phenomenon reflects that the source audio and video representations have high redundancy. The calculation of inter and intra relations becomes complicated without reducing redundancy. The complementarity of source audio and video representation can not be well expressed. The second phenomenon is that the accuracy of CapsGCN gets a great improvement. This situation fully reflects the complementarity of multimodal capsules is well released. At the same time, the calculation of inter and intra relations affects the sensitivity to some emotions, such as Fear. The third phenomenon is CapsGCN_att achieves the best performance. The introduced attention mechanism alleviates the excessive influence of certain factors on fusion such as Fear, and Surprise which are recognized as the most difficult. Also, due to the addition of the attention weights, the redundancy is further reduced.



	An	Di	Fe	Ha	Sa	Su
An	0.75	0.13	0.05	0.03	0.05	0.00
Di	0.05	0.93	0.03	0.00	0.00	0.00
Fe	0.07	0.07	0.68	0.00	0.15	0.03
Ha	0.10	0.10	0.03	0.78	0.00	0.00
Sa	0.00	0.07	0.07	0.00	0.85	0.00
Su	0.10	0.15	0.07	0.17	0.10	0.40

(a) CapsNet

	An	Di	Fe	Ha	Sa	Su
An	0.78	0.07	0.00	0.03	0.13	0.00
Di	0.00	0.95	0.00	0.00	0.05	0.00
Fe	0.13	0.10	0.33	0.00	0.42	0.03
Ha	0.07	0.10	0.00	0.78	0.03	0.03
Sa	0.00	0.03	0.00	0.00	0.97	0.00
Su	0.05	0.15	0.03	0.25	0.30	0.23

(b) GCN

	An	Di	Fe	Ha	Sa	Su
An	0.93	0.03	0.00	0.03	0.03	0.00
Di	0.00	0.97	0.00	0.00	0.03	0.00
Fe	0.05	0.10	0.63	0.00	0.15	0.07
Ha	0.03	0.10	0.00	0.88	0.00	0.00
Sa	0.00	0.03	0.07	0.00	0.90	0.00
Su	0.05	0.05	0.05	0.25	0.20	0.40

(c) CapsGCN

	An	Di	Fe	Ha	Sa	Su
An	0.88	0.03	0.03	0.07	0.00	0.00
Di	0.00	0.93	0.05	0.00	0.03	0.00
Fe	0.03	0.05	0.70	0.00	0.10	0.13
Ha	0.03	0.05	0.00	0.90	0.00	0.03
Sa	0.00	0.03	0.05	0.00	0.90	0.03
Su	0.13	0.03	0.00	0.20	0.10	0.55

(d) CapsGCN_att

Fig. 5. The confusion matrices.

4. CONCLUSION

In this paper, we studied the importance of information redundancy and information complementarity in multimodal fusion methods for MER. The effectiveness of the proposed fusion method CapsGCN has been verified under comparative experiments and ablation studies on eNTERFACE05'. Compared with the traditional fusion method, the classification accuracies achieve 80.83% with absolute increments more than 11.66% and 9.16%. The proposed method shows high sensitivity to all six emotions, especially for Fear, Happiness, and Suprise. The proposed fusion method also shows great potential to learn and model textual information. In the future, we plan to investigate the performance of the proposed model on some other multimodal datasets.

5. ACKNOWLEDGEMENTS

This work was supported by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

6. REFERENCES

[1] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] Yuanyuan Zhang, Zi-Rui Wang, and Jun Du, "Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–8.

[3] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7174–7178.

[4] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Yuan Gao, Lili Guo, and Jianwu Dang, "Temporal attention convolutional network for speech emotion recognition with latent representation," *Proc. Interspeech 2020*, pp. 2337–2341, 2020.

[5] Jinting Wu, Yujia Zhang, and Xiaoguang Zhao, "A generalized zero-shot framework for emotion recognition from body gestures," *arXiv preprint arXiv:2010.06362*, 2020.

[6] Abdulrahman Alreshidi and Mohib Ullah, "Facial emotion recognition using hybrid features," in *Informatics*. Multidisciplinary Digital Publishing Institute, 2020, vol. 7, p. 6.

[7] Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran, "Transformer based deep intelligent contextual embedding for twitter sentiment analysis," *Future Generation Computer Systems*, vol. 113, pp. 58–69, 2020.

[8] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[9] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[10] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Dan Ou, and Bo Zheng, "Adversarial multimodal representation learning for click-through rate prediction," in *Proceedings of The Web Conference 2020*, 2020, pp. 827–836.

[11] Ayush Kumar and Jithendra Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4477–4481.

[12] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[13] Bagus Tris Atmaja and Masato Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4482–4486.

[14] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3507–3511.

[15] S. Sabour, N. Frosst, and G.E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[16] K. Duarte, Y.S. Rawat, and M. Shah, "Videocapsulenet: A simplified network for action detection," in *Advances in Neural Information Processing Systems*, 2018, pp. 7610–7619.

[17] B.W. Zhang, X.F. Xu, M. Yang, X.J. Chen, and Y.M. Ye, "Cross-domain sentiment classification by capsule network with semantic rules," *IEEE Access*, vol. 6, pp. 1–1, 2018.

[18] Y. Min, M. Zhao, J.B. Ye, Z.Y. Lei, Zhao Z, and S. Zhang, "Investigating capsule networks with dynamic routing for text classification," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3110–3119, 2018.

[19] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling, "Modeling relational data with graph convolutional networks," in *European Semantic Web Conference*. Springer, 2018, pp. 593–607.

[20] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[21] Xinyu Zhou, Peifeng Li, Qiaoming Zhu, and Fang Kong, "Incorporating temporal cues and ac-gcn to improve temporal relation classification," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2020, pp. 580–592.

[22] Pinlong Zhao, Linlin Hou, and Ou Wu, "Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification," *Knowledge-Based Systems*, vol. 193, pp. 105443, 2020.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, and I. Kaiser, L.and Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[25] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *International Conference on Data Engineering Workshops*, 2006.

[26] Jiaxing Liu, Zhilei Liu, Longbiao Wang, Lili Guo, and Jianwu Dang, "Time-frequency deep representation learning for speech emotion recognition integrating self-attention," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 681–689.

[27] Laurens Van Der Maaten, "Learning a parametric embedding by preserving local structure," *Journal of Machine Learning Research*, vol. 5, pp. 384–391, 2009.