entropy

MDPI

Article

LGCCT: A Light Gated and Crossed Complementation Transformer for Multimodal Speech Emotion Recognition

Feng Liu [1,2,3,†], Si-Yuan Shen [2,†], Zi-Wang Fu [3], Han-Yang Wang [2], Ai-Min Zhou [1,2,4,*] and Jia-Yin Qi [5,*]

1 Institute of AI for Education, East China Normal University, Shanghai 200062, China; 52205901024@stu.ecnu.edu.cn
2 School of Computer Science and Technology, East China Normal University, Shanghai 200062, China; 51215901045@stu.ecnu.edu.cn (S.-Y.S.); 51215901028@stu.ecnu.edu.cn (H.-Y.W.)
3 School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China; fuziwang@bupt.edu.cn
4 Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention, School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China
5 Institute of Artificial Intelligence and Change Management, Shanghai University of International Business and Economics, Shanghai 200062, China
* Correspondence: amzhou@cs.ecnu.edu.cn (A.-M.Z.); ai@suibe.edu.cn (J.-Y.Q.)
† These authors contributed equally to this work.

Abstract: Semantic-rich speech emotion recognition has a high degree of popularity in a range of areas. Speech emotion recognition aims to recognize human emotional states from utterances containing both acoustic and linguistic information. Since both textual and audio patterns play essential roles in speech emotion recognition (SER) tasks, various works have proposed novel modality fusing methods to exploit text and audio signals effectively. However, most of the high performance of existing models is dependent on a great number of learnable parameters, and they can only work well on data with fixed length. Therefore, minimizing computational overhead and improving generalization to unseen data with various lengths while maintaining a certain level of recognition accuracy is an urgent application problem. In this paper, we propose LGCCT, a light gated and crossed complementation transformer for multimodal speech emotion recognition. First, our model is capable of fusing modality information efficiently. Specifically, the acoustic features are extracted by CNN-BiLSTM while the textual features are extracted by BiLSTM. The modality-fused representation is then generated by the cross-attention module. We apply the gate-control mechanism to achieve the balanced integration of the original modality representation and the modality-fused representation. Second, the degree of attention focus can be considered, as the uncertainty and the entropy of the same token should converge to the same value independent of the length. To improve the generalization of the model to various testing-sequence lengths, we adopt the length-scaled dot product to calculate the attention score, which can be interpreted from a theoretical view of entropy. The operation of the length-scaled dot product is cheap but effective. Experiments are conducted on the benchmark dataset CMU-MOSEI. Compared to the baseline models, our model achieves an 81.0% F1 score with only 0.432 M parameters, showing an improvement in the balance between performance and the number of parameters. Moreover, the ablation study signifies the effectiveness of our model and its scalability to various input-sequence lengths, wherein the relative improvement is almost 20% of the baseline without a length-scaled dot product.

Keywords: entropy invariance; multimodal speech emotion recognition; cross-attention; gate control; lightweight model; computational affection

Citation: Liu, F.; Shen, S.-Y.; Fu, Z.-W.; Wang, H.-Y.; Zhou, A.-M.; Qi, J.-Y. LGCCT: A Light Gated and Crossed Complementation Transformer for Multimodal Speech Emotion Recognition. *Entropy* **2022**, *24*, 1010. https://doi.org/10.3390/e24071010

Academic Editors: Qiang Zhang and Yifeng Zeng

Received: 7 July 2022
Accepted: 19 July 2022
Published: 21 July 2022

1. Introduction

Emotion plays a key role in interpersonal communication [1], wherein not only linguistic messages but also acoustic messages convey individual emotional states. In

many areas, such as human–computer interaction (HCI) [2], healthcare and cognitive sciences, much emphasis has been placed on developing tools to recognize human emotion in vocal expressions [3]. Recent booming advances in deep learning also promote the development of emotion recognition [4], a research field enabling machines to identify human emotion. Meanwhile, application requirements push the progress of lightweight models with high performance.

Focusing on acoustic features, a number of works have contributed to improving the performance of speech emotion recognition. Based on low-level descriptors (LLDs) in short frames, acoustic feature representations are extracted by deep learning networks, such as a convolutional neural network (CNN) [5], recurrent neural network (RNN) [6], etc. Some variant module architectures like CNN-LSTM [7], have also been developed to extract feature sequences and capture temporal dependencies.

Undoubtedly, linguistic information and acoustic information matters to speech emotion recognition [8]. Thus, both textual modality and audio modality should be taken into account to accomplish the task of multimodal emotion recognition. For audio modality, the process of feature extraction resembles that in unimodal speech emotion recognition. For textual modality, word-embedding models like GloVe [9] are commonly utilized.

What makes multimodal emotion recognition more challenging than unimodal emotion recognition is the process of modality fusion. Some early works concatenate different features as the input to the deep neural network [10]. To fuse the modalities in a deeper level, the standard transformer architectures [11] are widely extended to aggregate knowledge from one modality to the other, and, in this way, the learned modality-fused representation is enhanced [12,13].

Notwithstanding improvements made by prior works, the proportion of the modality-fused representation is seldom considered. To tackle this problem, we apply the gate-control mechanism [14] to the cross-attention module to decide whether to keep the source modality information or override the target modality information.

Most of the high performance of existing models are dependent on a great number of learnable parameters [15,16], ignoring the potential application in some promising areas like human–computer interaction (HCI), which requires real-time and light models. Thus, a lightweight model is necessary to improve the feasibility and practicability of the application of speech emotion recognition. Additionally, the transformer may have difficulty generalizing to a sequence with a different length than the fixed ones while training, which impair the performance under actual HCI scenarios. To handle this problem, we follow prior works [17], just multiplying attention logits by a hyperparameter, and justify it from a view of entropy.

In this paper, we propose LGCCT, a lightweight gated and crossed complementation transformer for multimodal speech emotion recognition. First, we utilize CNN-BiLSTM and BiLSTM [18] to extract audio features and textual features, respectively. Then the cross-attention module reinforces one modality feature with the other, and the gate mechanism functions as a flow control unit to balance the proportion of the two modalities and the length-scaled dot-product operation enhance the generalization to unseen sequence length. Finally, the fully connected layers followed by the transformer encoder layers predict the emotion.

Our contribution can be summarized as follows.

- We propose a model to fuse the text-modality and audio-modality representation and learn the mapping from modality-fused representation to emotion categories.
- We adopt length-scaled attention module to improve the performance of the model when applied to various testing sequence length and theoretically interpret the determination of the scaled hyperparameter from a view of entropy.
- We apply a gate-control mechanism to the traditional cross-attention module. The effectiveness is verified by the ablation study.
- We reach a balance between the performance and the number of parameters (only 0.432M). Experiments are conducted on the CMU-MOSEI dataset [19]. The experi-

ments also prove the generalization of our model to unseen sequence length. Compared with the baseline without a length-scaled dot product, the relative improvement is about 20%.

2. Related Works

Some early works for unimodal speech emotion recognition use traditional machine learning methods, such as a hidden Markov model [20], decision tree [21], and support vector machines [22]. With the development of deep learning methods, deep neural network (DNN)-based models in speech emotion recognition have thrived, like convolutional neural networks (CNN), recurrent neural networks (RNN) and long-short-term memory (LSTM) networks [6,7]. Some early works construct utterance-level features from segment-level probability distributions, and the extreme learning machine learns to identify utterance-level emotions [23]. Ref. [24] proposes a DNN-decision tree SVM model to extract the bottleneck features from confusion degree of emotion. CNNs mostly use spectrograms or audio features such as mel-frequency cepstral coefficients (MFCCs) and low-level descriptors (LLDs) as the inputs, followed by fully connected layers to predict the emotions [25]. RNN- and LSTM-based models take the temporal features into consideration and tackle this problem through sequence modeling [26]. Hybrid models like CNN-BiLSTM have also been adopted to effectively learn the information that represents emotions directly from conventional audio features [7,27]. Recently, the attention-based models and transformers have made significant progress in a range of fields [28,29]. Attention modules are used to learn the short-time frame-level acoustic features that are emotionally relevant, so that the temporal aggregated features can serve as more discriminative representation for classification [6]. Ref. [28] incorporates muti-task learning with attention-based hybrid models to better represent emotion features.

Emotion recognition in natural language processing (NLP) is also called sentiment analysis [30]. Early works take as input word embeddings, such as GloVe [9] and word2vec [31]. RNNs are capable of encoding the relations between sentences and capturing semantic meaning to distinguish sentiment better [32]. TextCNN [33] is a well-known convolutional neural network for sentence classification and is also widely applied to sentiment analysis [34]. The idea of attention is also popular in NLP. Ref. [21] uses a 2-D matrix to represent the embedding and introduces self-attention to extract an interpretable sentence embedding. In recent years, transformer-based self-supervised pretrained models, like BERT, thrive in NLP [11,35]. An increasing number of works take pretrained models as an encoder and get great performance boost [36,37].

Considering the fact that audios are composed of not only speech but also textual content, which can be extracted from the audio-based data, multimodal approaches using acoustic and lexical knowledge have also been explored. To further improve the accuracy, approaches that fuse audio, video and text are also a hot topic. There are mainly three kinds of future fusion strategies: attention-based feature fusion, GNN-based feature fusion [38–40] and loss-function-based feature fusion. For attention-based strategies, Ref. [18] proposes the bi-bimodal fusion network (BBFN) that performs fusion and separation on pairwise modality representations. Ref. [41] combines multi-scale CNN, statistical pooling unit and an attention module to exploit both acoustic and lexical information from speech. Ref. [13] proposes a multimodal transformer with the cross-modal attention mechanism to address this problem in an end-to-end manner. With such an idea, Ref. [42] uses both cross-modal attention and self-attention to propagate information within each modality. Ref. [43] designs a novel sparse transformer block to relieve the computational burden. Refs. [44,45] do the feature-fusion task by transferring it to a bi-modal translation task. For GNN-based strategies, Ref. [46] uses GCN to explore a more effective way of utilizing both multimodal and long-distance contextual information. For loss-function-based strategies, Ref. [47] hierarchically maximizes the mutual information in unimodal input pairs and between the multimodal fusion result and the unimodal input in order to maintain task-related information through multimodal fusion.

However, these methods ignore the fact that speech emotion recognition is needed mostly for real-time applications. Besides improving the accuracy by stacking models and arithmetic power, other factors such as being lightweight and showing efficiency and scalability to testing sequences with unfixed lengths are also necessary for practical applications. Thus, we will focus on reducing the number of parameters and improving the generalization to different testing sequences while maintaining performance.

3. Methodology

In this section, we will introduce the architecture of our network shown in Figure 1. The audio sequences are encoded by CNN-BiLSTM, while the text sequences are encoded by BiLSTM. Then the cross-attention module is utilized to fuse one modality representation into another modality representation respectively. The integration of the original-modality and the fused-modality representation is then controlled by the retain gate and the compound gate. The length-scaled coefficient is introduced while calculating the attention matrix to improve the generalization to different lengths, the validity of which will be illustrated from a view of entropy. To enhance the feature representation, the stacked transformer encoder is followed, and the classification is completed by the fully-connected layers. We will then illustrate our model in detail.

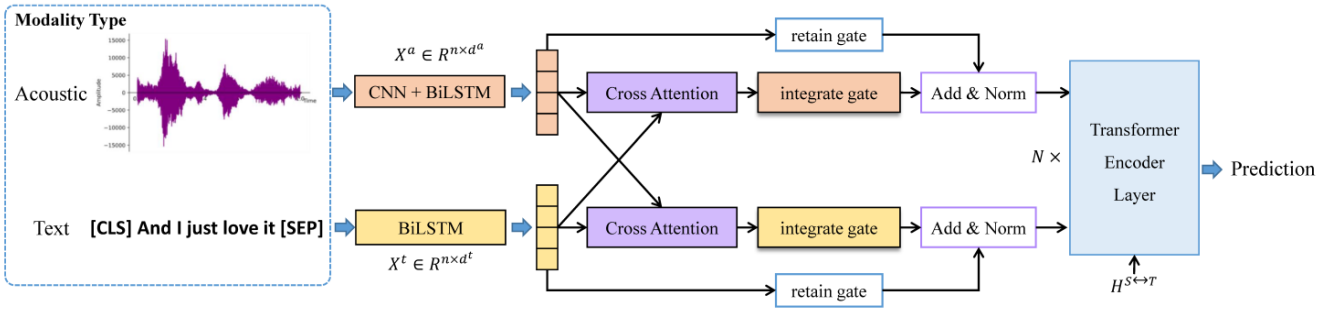

Figure 1. The overall architecture of LGCCT. CNN−BiLSTM and BiLSTM extract acoustic features and text features respectively. At the heart of the model, the cross−attention module with a gate−control mechanism fuses the modality information. The transformer encoder layers reinforce the modality-fused representation.

3.1. Text Encoder

We denote the text sequence as $X^t = \{x_1^t, x_2^t, \cdots, x_n^t\} \in \mathbb{R}^{n\times d^t}$, where d^t denotes the word embedding dimension and n denotes the length of the sequence. $BiLSTM$ is applied to capture the textual contextual dependencies.

$$H^t = BiLSTM(X^t) \tag{1}$$

where $H^t = \{h_1^t, h_2^t, \cdots, h_n^t\} \in \mathbb{R}^{n\times 2d^{t\prime}}$; H^t is the encoded feature representation, and $d^{t\prime}$ is the dimension of the hidden states.

3.2. Audio Encoder

We denote the audio sequence as $X^a = \{x_1^a, x_2^a, \cdots, x_n^a\} \in \mathbb{R}^{n\times d^a}$, where d^a denotes the dimension of low-level acoustic features and n denotes the length of the sequence. For convenience, we set the length of the audio sequence equal to that of the text sequence (namely, d^a is equal to d^t). Convolution layers are designed to extract high-level feature representation. Specifically, we use $Conv1d$ to integrate the temporal information.

$$H_{CNN}^a = Conv1d(X^a) \tag{2}$$

Then the $BiLSTM$ takes H^a_{cnn} as input and outputs the audio contextual feature representation.

$$H^a = BiLSTM(H^a_{CNN}) \tag{3}$$

where $H^a = \{h^a_1, h^a_2, \cdots, h^a_n\} \in \mathbb{R}^{n \times 2d^{a\prime}}$ is the encoded feature representation, and $d^{a\prime}$ is the dimension of the hidden states.

3.3. Cross-Attention Module

As shown in Figure 1, we use the transformer encoder to generate modality-fused representation so that the two kinds of modality information are fused bidirectionally. Herein, we define the source-modality representation as $H^S \in \mathbb{R}^{n \times d^S}$ and the target modality representation as $H^T \in \mathbb{R}^{n \times d^T}$, where $\{S, T\} \in \{t, a\}$. The process of modality fusion can be formulated as follows.

$$Q = W_Q \times H^T \tag{4}$$

$$K = W_K \times H^S \tag{5}$$

$$V = W_V \times H^S \tag{6}$$

where $Q \in \mathbb{R}^{n \times d^Q}$ is the query matrix, $K \in \mathbb{R}^{n \times d^K}$ is the key matrix' $V \in \mathbb{R}^{n \times d^V}$ is the value matrix, and $\times$ denotes matrix multiplication. Specifically, we set d^Q, d^K, d^V equal to the dimension of target modality d^T, denoted as d in the following. The source modality is transformed to the pair of key and value information while the target modality is transformed into the query information.

Then the fused-modality representation $H' \in \mathbb{R}^{n \times d^V}$ is calculated by length-scaled dot-product attention.

$$A = softmax\left(\frac{\lambda Q K^T}{\sqrt{\mathrm{d}}}\right) \tag{7}$$

$$H' = \mathrm{A}V \tag{8}$$

where $softmax$ operation is applied to the dimension of sequence; $A \in \mathbb{R}^{n \times n}$ is the attention matrix, and λ is a hyperparameter to enable well length generalization, which will be illustrated in the next section.

Following the cross-attention module, layer normalization is designed to attend to original modality in the other modality.

$$h^{S \to T} = LN(H' + H^T) \tag{9}$$

where LN means layer normalization.

To aggregate the feature representation, a fully connected feed-forward network is utilized after the cross-attention module.

$$H^{S \to T} = LN\left(h^{S \to T} + FFN(h^{S \to T})\right) \tag{10}$$

where FFN means fully connected feed-forward network. The overview of cross-attention module is provided in Figure 2.

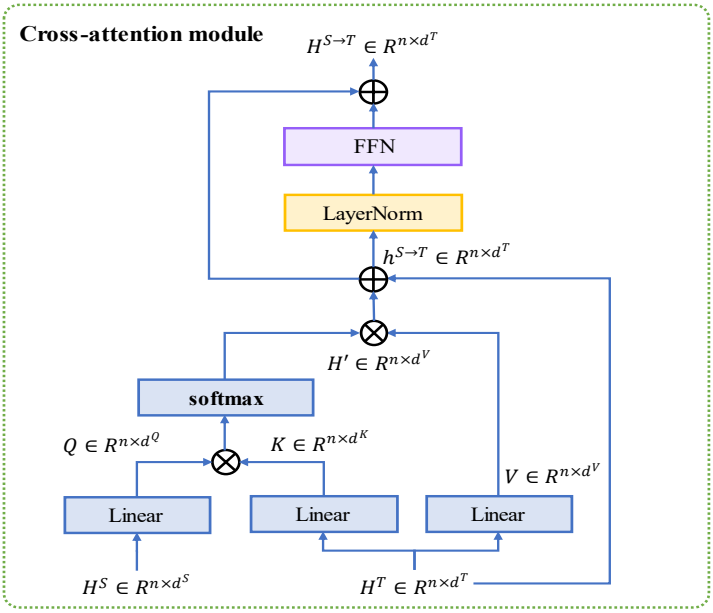

Figure 2. Cross-attention module fuses the modality information.

3.4. Entropy Invariance for Attention Operation

Following [17,48], we introduce a constant λ to improve the length generalization of attention operation, which can be interpreted from the perspective of entropy. In real-world applications, the length of an input sequence can be arbitrary, although the length is fixed during the training phase. The attention weight of the same token calculated in Equation (7) is supposed to converge to the same value independent of the length. From the view of entropy, we can consider the uncertainty as the degree of attention focus and revisit Equation (7) as follows.

$$p_{ij} = \frac{e^{\lambda q_i \cdot k_j}}{\sum_{j=1}^{n} e^{\lambda q_i \cdot k_j}} \tag{11}$$

$$\mathcal{H}_i = -\sum_{j=1}^{n} p_{ij} \log p_{ij} \tag{12}$$

where $q_i \in \mathbb{R}^d$ is the i-th query vector in the input sequence; $k_j \in \mathbb{R}^d$ is the j-th key vector. $q_i \cdot k_j$ is the dot product of these two vectors, reflecting the similarity of the two vectors, and p_{ij} is the attention score between the i-th token and j-th token in the sequence with total length of n. It should be noted that p_{ij} is actually the element at row i and column in the attention matrix A. $\mathcal{H}_i$ is the entropy of the i-th token.

We can take the attention score as uncertainty. Specifically, the entropy is zero when the attention is attended to only one token, and the entropy is $\log n$ when the attention is distributed uniformly. Then we provide an approximate theoretical justification for the determination of hyperparameter λ. Equation (12) can be rewritten as:

$$\mathcal{H}_i = \log \sum_{j=1}^{n} e^{\lambda q_i \cdot k_j} - \lambda \sum_{j=1}^{n} p_{ij} e^{\lambda q_i \cdot k_j} = \log n + \log \tfrac{1}{n} \sum_{j=1}^{n} e^{\lambda q_i \cdot k_j} - \lambda \sum_{j=1}^{n} p_{ij} e^{\lambda q_i \cdot k_j} \tag{13}$$

Here the second term can be approximated as:

$$\log \frac{1}{n} \sum_{j=1}^{n} e^{\lambda q_i \cdot k_j} \approx \log \exp \left(\frac{1}{n} \sum_{j=1}^{n} \lambda q_i \cdot k_j \right) = \log \lambda \overline{q_i \cdot k_j} \tag{14}$$

Based on the hypothesis that the softmax operation can be used as a continuous, differentiable approximation to argmax [49], the third term can be approximated as:

$$\lambda \sum_{j=1}^{n} p_{ij} e^{\lambda q_i \cdot k_j} \approx \lambda \max_{1 \le j \le n} (e^{\lambda q_i \cdot k_j}) \tag{15}$$

Therefore, we have:

$$\mathcal{H}_i \approx \log n - \lambda \left(\max_{1 \le j \le n} (e^{\lambda q_i \cdot k_j}) - \log \lambda \overline{q_i \cdot k_j} \right) \tag{16}$$

To mitigate the influence of the sequence length n, $\lambda \propto \log n$. For convenience, we set λ as $\log n$. Equation (7) can be redefined as follows:

$$A = softmax\left(\frac{\log n Q K^T}{\sqrt{\mathrm{d}}} \right) \tag{17}$$

In this way, the contribution of the input token is more stable so that the attention matrix is theoretically more robust to the variation of input length. We determine the value of λ, and the next procedure is the same as mentioned above.

3.5. Gate Control

Based on the idea of close and open access of information flow [14], the gate unit is introduced to our network architecture. Some of the original-modality information should be attended to the fused-modality information.

$$H^{S \to T} = H^{S \to T} \times G_i + H^T \times G_r \tag{18}$$

where $G_i \in \mathbb{R}^{d \times d}$ represents the learnable integrate gate, and $G_r \in \mathbb{R}^{d \times d}$ represents the learnable retain gate. With learnable weights, the integrate gate decides how much fused information should be combined, and the retain gate decides how much original information should be preserved.

3.6. Classification

The transformer encoder layer is then employed, taking the concatenation of bidirectional modality information as input, as shown in Figure 3. Eventually, the classification is performed by fully connected feed-forward network.

$$\widehat{X} = FFN\Big(Transformer([H^{S \to T}, H^{T \to S}]) \Big) \tag{19}$$

where $[\cdot,\cdot]$ denotes the concatenation of bidirectional modality information; transformer denotes the transformer encoder layer, and $\widehat{X}$ denotes the predicted emotion category.

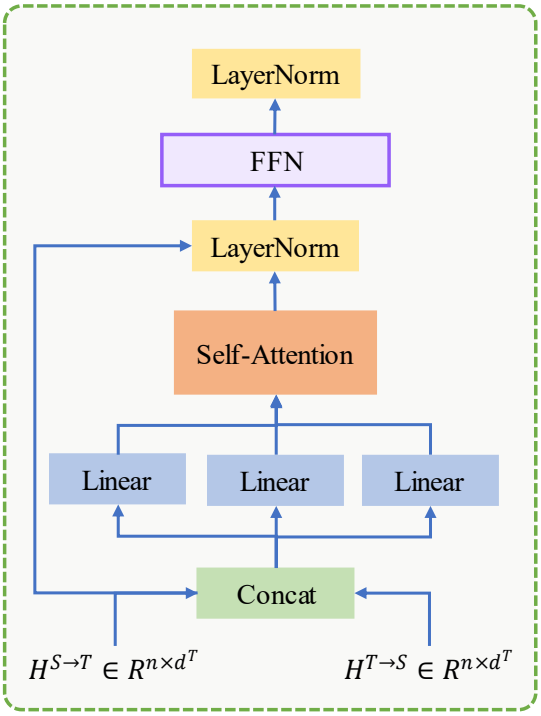

Figure 3. The architecture of the Transformer.

4. Results

In this section, we evaluate our model on CMU-MOSEI [19]. The implementation details and the experimental results are illustrated in this part.

4.1. Dataset and Metrics

CMU-MOSEI is a human multimodal sentiment-analysis dataset consisting of 23,453 sentences from YouTube videos, involving 1000 distinct speakers and 250 topics. The gender in the dataset is balanced (57% male to 43% female). The average length of sentences is 7.28 s, and acoustic features are extracted at a sampling rate of 20 Hz. Each sample is labeled by human annotators with a sentiment score from -3 to 3, including highly negative, negative, weakly negative, neutral, weakly positive, positive and highly positive.

The train/validation/test splits are provided by the CMU Multimodal Data SDK, wherein the same speaker does not appear in both train and test splits to ensure speaker independency. The length of the aligned sequences is 50. Using P2FA [50], the audio stream is aligned with the word along the timestep, within which the two modality features are averaged. For the text modality, the transcripts are processed with pre-trained GloVe [9] word embeddings, and the embeddings are 300-dimensional vectors. For the audio modality, the low-level 74-dimension vectors are extracted by COVAREP [51], including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients.

Consistent with the previous works [12,52], we adopt the metrics of 7-class accuracy (from strongly negative to strongly positive), binary accuracy (positive/negative sentiments) and F1 score (harmonic mean of the binary precision and recall). Specifically, the predicted digit will be rounded first. For 7-class accuracy, the predicted digit will be compared with the annotated sentiment score from -3 to 3. For binary accuracy, the predicted digit will be classified to positive or negative sentiment according to its positivity and negativity. The F1 score is the harmonic mean of the binary precision and recall.

4.2. Implementation Details

Our LGCCT is implemented by Pytorch [53] and optimized by Adam [54], with a learning rate 1×10^{-3}, 40 epochs, a decay rate of 0.1 after 20 epochs, batch size of 24, a

gradient clip of 1.0 and an output dropout rate of 0.1. The other hyperparameter mentioned in Section 3 are shown in Table 1. The dimensions of the hidden states H in the transformer are unified to 30.

Table 1. Detailed dimensions of LGCCT.

Notation	Meaning	Value
n	Aligned input-sequence length	50
d^t	Word-embedding dimension	300
d^a	Audio feature dimension	74
$d^{t\prime}$	Encoded text feature dimension by BiLSTM	32
$d^{a\prime}$	Encoded audio feature dimension by CNN-BiLSTM	32
d	Hidden state dimension	30
λ	Length-scale logits	$\log 50$

The hardware environment for running is as follows: CPU: Intel(R) Xeon(R) Silver 4210R @ 2.40 GHz; GPU: NVIDIA Quadro RTX 8000; system running environment: Ubuntu 18.04.6.

4.3. Baselines

EF-LSTM. Early fusion LSTM concatenates the inputs from different modalities as the input to a single LSTM and classifies the feature vectors.

LF-LSTM. Late fusion LSTM describes each modality information separately, and the fusion takes place at the decision level.

RAVEN [55]. The proposed recurrent attended variation embedding network is composed of three parts, including nonverbal subnetworks, gated modality-mixing network and multimodal shifting.

MCTN [56]. The cyclic translation mechanism based on RNN is designed to learn joint representations.

MulT [12]. This model uses the cross-modal transformer, namely a deep stack of several cross-modal attention blocks, to fuse different modalities.

MISA [16]. Based on LSTM and pretrained BERT [35], MISA projects each modality to two subspaces.

BBFN [15]. The BERT encoder is utilized to obtain feature representation, which is then fused by transformer-like modules.

4.4. Comparison with Baseline Models

As shown in Table 2 and Figure 4, our model shows superiority in the balance between parameters and performance and maintains the higher F1 score than almost all models with limited number of parameters. Acc_7 denotes 7-class accuracy; Acc_2 denotes binary accuracy, and $F1$ denotes the $F1$ score. In terms of performance, our model outperforms the other models in Acc_2 and $F1$ score, except for BBFN, which utilizes the BERT encoder, requiring a large number of trained parameters and memory space. It is noteworthy that the axis of parameters in Figure 4 omits the range from 1.4 M and 110.4 M, and, thus, the required parameters for BBFN and MISA are huge. Considering the balance between parameters and performance, the model scale is enlarged by more than 110 M parameters for about 4% absolute performance improvement, indicating the imbalance tradeoff between performance and computational complexity. Besides, in almost all metrics, our model performs better than the three models trained with tri-modality information, learning richer modality information. In terms of the number of parameters, our model ranks only second to MCTN. However, the Acc_2 and $F1$ scores of our model are much higher than that of MCTN, by more than 2%, while the Acc_7 is slightly worse than that of MCTN by 0.14%. Although the parameters of LSTM-based models are low, performance is also limited, and our LGCCT surpass them on almost all of the metrics. This is partly due to the fact that they do not take into account the interactions between the modalities, just concatenating the modalities. Our model adopts efficient modules, such as cross-attention and a gate-control mechanism, to fuse the modality information and maintain the

balance between the source modality and target modality. The experiments indicate the effectiveness of our model.

Table 2. The performance and the number of parameters on the CMU-MOSEI dataset.

Method	#Params(M)	Acc_7(%)	F_1(%)	Acc_2(%)
MulT	0.961	48.2	80.2	79.7
MCTN	0.247	47.64	78.87	77.86
MISA **	110.915	53.31	80.81	80.26
BBFN **	110.548	51.7	85.5	85.5
EF-LSTM *	0.56	47.4	78.2	77.9
LF-LSTM *	1.22	48.8	80.6	80.6
RAVEN *	1.19	45.5	75.4	75.7
LGCCT (ours)	0.432	47.5	81.0	81.1

* with tri-modality, namely audio, video and text. ** with pretrained BERT.

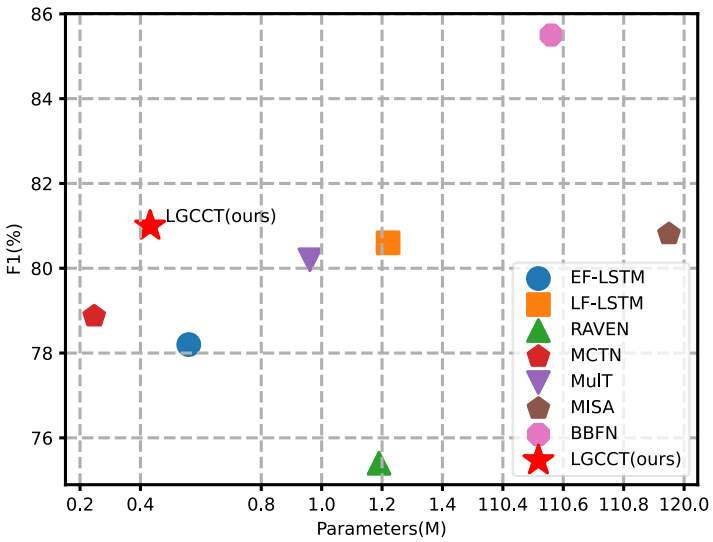

Figure 4. Comparison of the F_1 score of different models on CMU-MOSEI. The proposed LGCCT achieves the best performance with an order of magnitude smaller model size.

4.5. Ablation Study

To study the effect of different parts on the performance, we conduct some experiments on the CMU-MOSEI dataset. The results are shown in Table 3. First, we evaluate the influence of the gate-control mechanism. The Acc_7 drops about 4.6%, indicating the effectiveness of the introduced gate units. Second, the audio encoder CNN-BiLSTM and text encoder BiLSTM are removed. The LGCCT model outperforms these two models in all metrics, suggesting the importance of feature extraction. It is noteworthy that the performance degrades most when the feature encoder is removed, signifying the fact that the feature encoder aggregates the original modality information and that the representation to the modality-fusing modules is powerful. Finally, the transformer encoder ahead of the fully connected layer is removed. The results show that it is necessary to apply the self-attention module to encode the modality-fused representation. However, the performance without a transformer degrades least, but the parameters are cut down most in the ablation study. We assume that the cross-attention operation in the modality-fusing module manages to attend to interactions between multimodal sequences, and, thus, the contribution of the last transformer to the performance is restricted. The ablation study implies the function of the components of our model and verifies the contribution of each module to the performance.

Table 3. Ablation study on the CMU-MOSEI dataset.

Model	#Params(M)	Acc_7(%)	Acc_2(%)	F_1(%)
LGCCT	0.432	47.5	81.0	81.1
w/o gates	0.429	42.9	76.7	76.3
w/o CNN-BiLSTM & BiLSTM	0.354	40.9	70.7	70.8
w/o Transformer	0.203	40.3	75.6	78.0

4.6. Length-Scaled Attention

To mitigate the problem that the length of the training sequence is fixed while the length of the testing sequence may vary, we introduce length-scaled attention. For the standard input, the length of the training sequence and the testing sequence are unified to 50, which is referred to in Tables 4 and 5. To validate the effectiveness of length-scaled attention, we clip the original sequence according to two proportions of the original length: 80% and 60% of the original one. This experiment configuration yields two other length: 40 and 30, respectively, denoted in Tables 4 and 5. Then we test/train the variant of LGCCT with/without length-scaled attention. Other configurations are kept default. Table 4 shows great improvement of the model when the length of the training sequence and the testing sequence varies, especially at a length of 30. To be specific, when training and testing on data with the same length, the effect of length-scaled attention is not obvious but length-scaling outperforms its counterpart by 12.9% and 6.9%, respectively, when training and testing on different lengths. A closer look at the result with a length of 30 shows that the model without length-scaled attention performs poorly when the testing sequence length is 50 but the training one is clipped to 30. For the binary accuracy in Table 4, the length-scaled dot product brings the relative improvement of 20%. A similar improvement is also shown in training all test settings with clipped length of 30, wherein the relative improvement is 11% on accuracy and 30% on the F1 score. Moreover, length scaling helps stabilize performance on the F1-score as shown in Table 5, while shorter testing sequences lead to serious performance degradation for vanilla attention operation, like a 57.8% F1 score when testing sequence is cut to 30. Interestingly, the model with length scaling does not show superiority over its counterpart without that, which to some extent reveals the data efficiency of our LGGCT when the length gap between the training and testing sequence is not large. We hypothesize such stable performance occurs because the length of text modality and audio modality is not always identical and is forced to be aligned, wherein sometimes zero logits are padded to the end of the sequence. This suggests that clipped data may compose of useless zero frames. Furthermore, the variant LGCCT with length-scaled attention manages to generalize to the sequence with a length different from the training set.

Table 4. Accuracy comparisons on CMU-MOSEI with different length distributions.

	All = 50	Part = 30		Part = 40	
Type	**Train All Test All**	**Train Part Test All**	**Train All Test Part**	**Train Part Test All**	**Train All Test Part**
Length scaled	80.8	75.7	65.2	74.4	67.7
w/o length scaling	81.1	62.8	58.3	77.0	71.9

Table 5. F1-score comparisons on CMU-MOSEI with different length distributions.

	All = 50	Part = 30		Part = 40	
Type	**Train All Test All**	**Train Part Test All**	**Train All Test Part**	**Train Part Test All**	**Train All Test Part**
Length scaled	80.7	76.2	75.3	76.8	74.8
w/o Length scaled	81.0	77.2	57.8	78.3	72.3

5. Conclusions

In this paper, we propose LGCCT, a lightweight gated and crossed complementation transformer for multimodal speech emotion recognition. Text encoder BiLSTM and audio encoder CNN-BiLSTM are utilized to obtain modality feature expression. At the heart of LGCCT, cross-attention modules fuse the modality information with each other, and the learnable gate-control mechanism controls the information flow and stabilizes the training process. Moreover, we apply length scaling to the attention module to improve the generalization of the transformer to various testing strings, which can be elaborated from a view of entropy invariance. In particular, the attention weight of the same token in the attention matrix is supposed to converge to the same value independent of the length. From the view of entropy, we can consider the uncertainty as the degree of attention focus. The attention scores can be consistent with the length of the input length just by multiplying the hyperparameter λ. At the top of the model, the fully connected forward network followed by the transformer encoder learns the mapping from modality-fused representation to emotion categories. In the experiments, we compare our model with baseline models on the benchmark dataset and further the ablation study. Our model achieves the balance between performance and the number of parameters with only 0.432 M parameters. The results also show the effectiveness of each component, underlying the performance and lightweight footprint of our model. Furthermore, the length-scale attention does help the model generalize to various sequence lengths under the experiment with different sequence lengths.

From the view of multimodal speech emotion recognition, our method has shown a balance between performance and the number of parameters. We attribute this to two factors. First, we adopt efficient modules in our network, such as cross-attention and a gate-control mechanism. In this way not only can the inter-modality information communicate and mingle with each other, but the generated modality-fused information can also maintain the balance between the target modality and the source modality. Similar efficiency is maintained within the process of feature extraction. Second, we keep the dimension of the hidden states low. Just as shown in Table 1, all of the dimension are kept approximately 30, except for the input channels. In contrast, many previous works adopting the BERT encoder [35] have to adjust the input channel to 512 [15,16], far more than ours. It is noteworthy that we do not apply any other down-sampling after the feature-extraction stage, keeping the embedding dimension low and the sequence length the same. This design is similar to the classic ideas in computer vision, namely maintaining large activation maps while decreasing the quantity of parameters [57]. For a wide neural network, not all of the information contained in high-dimensional vectors are useful [58], and thus our LGCCT is designed as a narrower network so that the information can be more compact.

From the view of information theory, our entropy-based LGCCT variant is capable of generalizing to testing sequences with various lengths. We consider the degree of attention focus as the uncertainty and let the same token converge to the same value independent of the length by multiplying the predefined constant λ. Since the attention matrix is computed by the learnable parameters, the model is supposed to learn the value ideally. However, our experiment shows that the model with length-scaling performs more stably, while the model without length scaling fluctuates in performance. This phenomenon indicates that the perspective of entropy really works, and the inductive bias [59] can help the model find better solutions. Actually, the idea of entropy is widely applied to deep learning methods. One of the most-typical applications is cross entropy. This criterion serves as a loss function to compute the loss between input and target, especially when handling a classification problem with multiple classes [60]. More generally, other classic perspectives in information theory have been used in deep learning methods [47,61,62].

In the future, we will design a more effective gate mechanism, following some gate units such as LSTM cell and GRU gates [63]. Furthermore, other modalities like video can be considered, so that a tri-modal emotion recognition network can be developed for application in realistic scenarios. The effectiveness of length-scaled attention for multimodal

emotion recognition may shed light on the wider usage of entropy, as well as information theory, in the deep learning community.

Author Contributions: Conceptualization, F.L., A.-M.Z. and J.-Y.Q.; methodology, F.L., S.-Y.S.; investigation, F.L., S.-Y.S. and Z.-W.F.; resources, F.L.; data curation, H.-Y.W. and Z.-W.F.; supervision, A.-M.Z. and J.-Y.Q.; project administration, A.-M.Z. and J.-Y.Q.; funding acquisition, A.-M.Z. and J.-Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by 2019, Digital Transformation in China and Germany: Strategies, Structures and Solutions for Ageing Societies, GZ 1570. Also supported by the Research Project of Shanghai Science and Technology Commission (No.20dz2260300) and The Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks to the blockchain technology team of the Cross-Innovation Laboratory of East China Normal University for their fully support. At the same time, We also thank the reviewers for their valuable time.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. The authors declare no conflict of interest.

References

1. Wang, X.; Chang, Y.; Sugumaran, V.; Luo, X.; Wang, P.; Zhang, H. Implicit Emotion Relationship Mining Based on Optimal and Majority Synthesis from Multimodal Data Prediction. *IEEE MultiMedia* **2021**, *28*, 96–105.
2. Card, S.K.; Moran, T.P.; Newell, A. *The Psychology of Human-Computer Interaction*; CRC Press: Boca Raton, FL, USA, 2018.
3. Lugović, S.; Dunđer, I.; Horvat, M. Techniques and applications of emotion recognition in speech. In Proceedings of the 2016 39th international convention on information and communication technology, electronics and microelectronics (mipro), Opatija, Croatia, 30 May–3 Jun 2016; pp. 1278–1283.
4. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80.
5. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-end speech emotion recognition using deep neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093.
6. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2227–2231.
7. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
8. Schuller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; pp. I–577.
9. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
10. Gu, Y.; Yang, K.; Fu, S.; Chen, S.; Li, X.; Marsic, I. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In Proceedings of the Conference Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; p. 2225.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
12. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Conference Association for Computational Linguistics, Florence, Italy, 28 July 2019–2 August 2019; p. 6558.
13. Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Niu, M. Multimodal transformer fusion for continuous emotion recognition. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3507–3511.

14. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
15. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-P.; Poria, S. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In Proceedings of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 6–15.
16. Hazarika, D.; Zimmermann, R.; Poria, S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1122–1131.
17. Chiang, D.; Cholak, P. Overcoming a Theoretical Limitation of Self-Attention. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 7654–7664.
18. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
19. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
20. Hu, H.; Xu, M.-X.; Wu, W. GMM supervector based SVM with spectral features for speech emotion recognition. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; pp. IV-413–IV-416.
21. Lin, Z.; Feng, M.; Santos, C.N.d.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. *arXiv* **2017**, arXiv:1703.03130.
22. Milton, A.; Roy, S.S.; Selvi, S.T. SVM scheme for speech emotion recognition using MFCC feature. *Int. J. Comput. Appl.* **2013**, *69*, 34–39.
23. Han, K.; Yu, D.; Tashev, I. Speech emotion recognition using deep neural network and extreme learning machine. In Proceedings of the INTERSPEECH, Singapore, 7–10 September 2014; pp. 223–227.
24. Sun, L.; Zou, B.; Fu, S.; Chen, J.; Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Commun.* **2019**, *115*, 29–37.
25. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; pp. 1–5.
26. Lee, J.; Tashev, I. High-level feature representation using recurrent neural network for speech emotion recognition. In Proceedings of the INTERSPEECH, Dresden, Germany, 6–10 September 2015; pp. 1–4.
27. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
28. Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2803–2807.
29. Wang, X.; Wang, M.; Qi, W.; Su, W.; Wang, X.; Zhou, H. A Novel end-to-end Speech Emotion Recognition Network with Stacked Transformer Layers. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6289–6293.
30. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. *arXiv* **2002**, arXiv:cs/0205070v1.
31. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
32. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
33. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.
34. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1253.
35. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
36. Huang, B.; Carley, K.M. Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5469–5477.
37. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 380–385.
38. Lian, Z.; Tao, J.; Liu, B.; Huang, J.; Yang, Z.; Li, R. Conversational Emotion Recognition Using Self-Attention Mechanisms and Graph Neural Networks. In Proceedings of the INTERSPEECH, Shangai, China, 25–29 October 2020; pp. 2347–2351.
39. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv* **2019**, arXiv:1908.11540.

40. Liu, J.; Chen, S.; Wang, L.; Liu, Z.; Fu, Y.; Guo, L.; Dang, J. Multimodal emotion recognition with capsule graph convolutional based representation fusion. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6339–6343.
41. Peng, Z.; Lu, Y.; Pan, S.; Liu, Y. Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3020–3024.
42. Sun, L.; Liu, B.; Tao, J.; Lian, Z. Multimodal Cross-and Self-Attention Network for Speech Emotion Recognition. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 4275–4279.
43. Cheng, J.; Fostiropoulos, I.; Boehm, B.; Soleymani, M. Multimodal Phased Transformer for Sentiment Analysis. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2447–2458.
44. Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; Kong, W. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 1–6 August 2021; pp. 5301–5311.
45. Wang, Z.; Wan, Z.; Wan, X. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 2514–2520.
46. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. *arXiv* **2021**, arXiv:2107.06779.
47. Han, W.; Chen, H.; Poria, S. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. *arXiv* **2021**, arXiv:2109.00412.
48. Su, J. Entropy Invariance in Softmax Operation. Available online: https://kexue.fm/archives/9034 (accessed on 11 April 2022).
49. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv* **2016**, arXiv:1611.01144.
50. Yuan, J.; Liberman, M. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* **2008**, *123*, 3878.
51. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; Scherer, S. COVAREP—A collaborative voice analysis repository for speech technologies. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 960–964.
52. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.-P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
53. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703.
54. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
55. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P.P.; Zadeh, A.; Morency, L.-P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 7216–7223.
56. Pham, H.; Liang, P.P.; Manzini, T.; Morency, L.-P.; Póczos, B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6892–6899.
57. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.
58. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
59. Baxter, J. A model of inductive bias learning. *J. Artif. Intell. Res.* **2000**, *12*, 149–198.
60. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
61. Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; Pérez, P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2517–2526.
62. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2018**, arXiv:1808.06670.
63. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.