KEY-SPARSE TRANSFORMER FOR MULTIMODAL SPEECH EMOTION RECOGNITION

Weidong Chen⋆   *Xiaofeng Xing*⋆   *Xiangmin Xu*⋆⋆   *Jichen Yang*⋆   *Jianxin Pang*†

⋆ School of Electronic and Information Engineering, South China University of Technology, China
† UBTECH Robotics Corp, China

ABSTRACT

Speech emotion recognition is a challenging research topic that plays a critical role in human-computer interaction. Multimodal inputs further improve the performance as more emotional information is used. However, existing studies learn all the information in the sample while only a small portion of it is about emotion. The redundant information will become noises and limit the system performance. In this paper, a key-sparse Transformer is proposed for efficient emotion recognition by focusing more on emotion related information. The proposed method is evaluated on the IEMOCAP and LSSED. Experimental results show that the proposed method achieves better performance than the state-of-the-art approaches.

Index Terms— speech emotion recognition, sparse network, modality interaction

1. INTRODUCTION

Speech emotion recognition (SER) is fast becoming a key instrument in human-computer interaction (HCI) [1]. SER also sheds new light on autism and the elderly care and so on, which are collectively referred to healthcare [2]. For example, the people who suffer from severe speech and language disorder have difficulty expressing their emotions. An emotion recognition system can help to treat the patients and improve their emotional communication skills.

Speech is multimodal as it contains text information by its nature. Latest researches [3, 4] have also proved that multimodal methods outperform the uni-modal methods. Consequently, multimodal SER has been a hot research topic in recent years. For example, *Yoon et al.* [5] use dual recurrent neural networks to combine the information from audio and text. In the same way, *Krishna et al.* [6] use raw audio waveform as audio features and GloVe word embeddings as text features for multimodal learning. Moreover, *Peri et al.* [7] combine audio and video information and utilize multitask setting for emotion recognition. In this paper, we use both audio and text information for SER.

Pre-trained Self Supervised Learning (SSL) has made great success in many fields such as natural language processing [8, 9] and speech recognition [10]. Meanwhile, recent
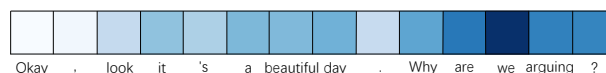
---
*Corresponding author. Email: xmxu@scut.edu.cn

Fig. 1. The attention weights of the utterance "OK, look it's a beautiful day. Why are we arguing?" in vanilla Transformer. Darker colors represent larger weights.

works [11, 12] that use SSL model have obtained promising results in SER. Nowadays, wav2vec [10] and RoBERTa [9] are the most commonly used pre-trained SSL models in the literature. Thus, in this paper, we use them to extract audio and text embeddings, respectively.

Inspired by the attention mechanism, Transformer [13], which is outstanding in modeling long sequence, is proposed and has achieved great success in natural language processing [11]. Meanwhile, several Transformer based architectures have been introduced for SER. *Tarantino et al.* [14] use global windowing system in Transformer to capture deep relationships within the utterance. Moreover, *Huang et al.* [15] use Transformer to fuse different modalities for sentiment analysis. In this paper, we use Transformer as our basic structure to implement emotion recognition.

However, few works have paid attention to that not all the information in audio or text is related to emotion. For example, considering a text *"Okay, look it's a beautiful day. Why are we arguing?"* in IEMOCAP [16], the attention weights in vanilla Transformer are shown in Figure 1. We can see that the attention weights in Transformer are assigned to all the words. However, words *"beautiful"* and *"arguing"* contain the majority of emotional information in this sentence. And the words that are not related to emotion such as *"it"*, *"a"* and *"look"*, are unnecessary for SER task and become noises, leading to the limitation of system performance. To address this issue, we propose a novel method, named key-sparse Transformer (KS-Transformer), to judge the importance of each word or speech frame in the sample and help the model focus more on the emotion related information. Based on KS-Transformer, we further design a cascaded cross-attention block to fuse different modalities with high efficiency.

The contributions of this paper can be summarized as follows:

- We propose KS-Transformer to judge the importance of each frame or word that helps the model focus more on the emotional information. Based on KS-Transformer, we further design a cascaded cross-attention block to achieve interaction between different modalities.

- We evaluate the proposed method on IEMOCAP and LSSED, and demonstrate that it achieves better results than the existing state-of-the-art approaches.

2. PROPOSED METHOD

The proposed model, as shown in Figure 2, mainly consists of three modules. In which, feature extraction module is used to learn the input features, modality interaction module is used for learning interactive information and deep fusion module aims to further combine the information from audio and text. Specifically, the first module (gray parts) is based on vanilla Transformer and the last two modules (yellow parts) are based on KS-Transformer. More details will be introduced in the following subsections.
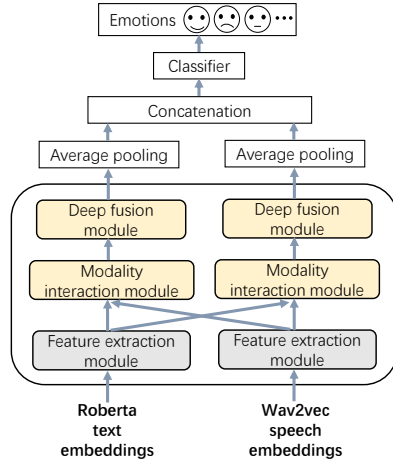

Fig. 2. Overview structure of the proposed model.

2.1. Key-Sparse Transformer

2.1.1. Vanilla Transformer

Vanilla Transformer consists of encoder and decoder originally. In this paper, we use Transformer to represent the encoder part, since it is the one needed for the implementation of our proposed architecture. The inputs of Transformer are divided into Q, K and V, which consist of Query, Key and Value vectors, respectively. The attention mechanism in vanilla Transformer is depicted as follows:

$$W = softmax(\frac{QK^T}{\sqrt{d_Q}}) \qquad (1)$$

$$attn = W \times V \qquad (2)$$

where d_Q is the dimension of the Query vector, W is the weight matrix and **attn** is the attention output. For multi-head attention mechanism, we combine the attention outputs from all the heads. More details can be found in [13].

2.1.2. Key-Sparse attention mechanism

The key-sparse Transformer aims to find the emotional information automatically. Assume the number of Query vectors in Q is i while that of Key vectors in K is j, the key-sparse attention mechanism is illustrated in Figure 3. It should be noted that K and V are always the same in Transformer.
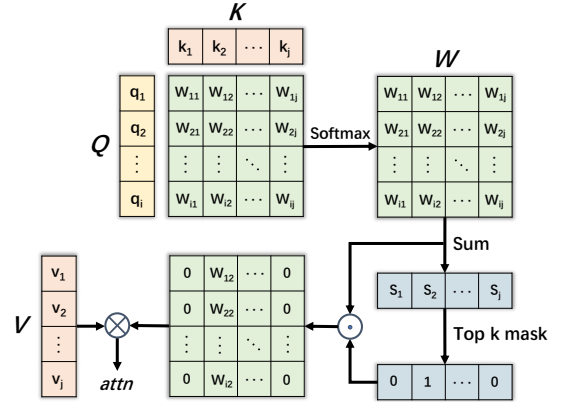

Fig. 3. The key-sparse attention in KS-Transformer. In which, softmax and summation are performed on each row and column, respectively. $\odot$ and $\otimes$ represent position-wise multiplication and matrix multiplication, respectively.

Key-sparse attention mechanism, which is used in KS-Transformer, is capable of judging the importance of each speech frame or word automatically. As shown in Figure 3, the weight matrix W is obtained by multiplying Q and K, and each row in W are the weights of Value vectors in V. As a Value vector represents a frame in audio or a word in text, we add up all the weights of the same Value vector and the summation is used as a discriminator for the importance of the speech frame or word in the sample. We select k Value vectors with *top-k* largest summation and keep their attention weights in weight matrix unchanged while the others are reset to zero. This operation makes the weight matrix from dense to sparse and reduces the redundancy, that's why we call the Transformer used here as KS-Transformer. The *top-k* mask is calculated by Equation 3.

$$M_z = \begin{cases} 0 & \text{if } s_z < threshold \\ 1 & \text{if } s_z \geq threshold \end{cases} \qquad (3)$$

where *threshold* is the kth largest summation and $z \in [1, j]$.

6898

2.2. Modality interaction module

Because modality interaction module is based on cascaded cross-attention block (CCAB), we introduce CCAB's structure first. As shown in the left part of Figure 4, CCAB is a cascade of two KS-Transformers, in which, the first KS-Transformer creates Q from modality A and K, V from modality B. With this special input method, the key-sparse attention mechanism will find out the most relevant part in B for A and produce an output which has combined A with B information. Since the emotional information between different modalities is often complementary [3, 17, 18], neither A nor B can represent the accurate emotion. Therefore, the second KS-Transformer in CCAB takes the fused features as input and considers the information from both modality A and modality B when applying key-sparse attention. Benefited from CCAB, A and B are fused more comprehensively and accurately.
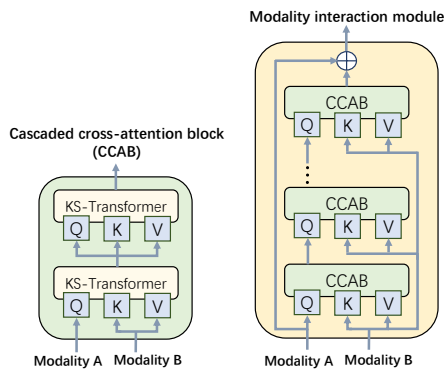

Fig. 4. The details of CCAB (left) and modality interaction module (right).

As shown in the right part of Figure 4, modality interaction module consists of a stack of CCABs, wherein the later CCAB takes the output of the former CCAB as Q input while K and V inputs are always from modality B. That the information from B goes through one CCAB is regarded as one interaction because the information from B had flowed into A by the key-sparse attention. More than one CCAB are applied for multiple times interactions. A skip connection is utilized for the features' stability.

2.3. Deep fusion module

Most researches take the fused features to predict emotions after the interaction [12, 19]. However, we argue that the fused features maybe not the best and can be deep fused to further improve the system performance. In detail, deep fusion module consists of several KS-Transformers, in which, they take the fused features as input and utilize key-sparse attention to enhance the interaction between audio and text and implement deep fusion.

3. EXPERIMENTS

3.1. Database introduction

IEMOCAP contains five sessions, every of which has one male and one female speaker, respectively. To stay consistent with the previous works [6, 17, 18], we use 5,531 utterances from four emotions: angry, neutral, happy (& excited) and sad. We conduct experiments in leave-one-session-out cross-validation strategy.

LSSED [20] is a new released large-scale English speech emotion dataset, which has data collected from 820 subjects and contains 147,025 samples. Consistent with [20], we use four emotion categories, including angry, neutral, happy and sad. For each emotion class, its associated samples are randomly split into train/development/test in ratio of 7/1/2, respectively. Every experiment is run for 10 times to avoid randomness, and the averaged result is used as the final accuracy.

3.2. Experimental setup

The pre-trained wav2vec and RoBERTa are available online[1]. The max lengths of the audio and text feature sequence are set to 460 and 20, respectively. SGD optimizer with a learning rate of 5×10^{-4} on IEMOCAP and 1×10^{-4} on LSSED is applied to optimize the model. The learning rate drops to 50% of the original every 30 epochs. Dropout with $p = 0.5$ is utilized to alleviate over-fitting. The batch size is 32.

Feature extraction module is used to learn the input features, which are extracted from pre-trained SSL models, aims to obtain suitable features for SER task. For modeling rich contexts, this module is based on vanilla Transformer. Q, K and V inputs here are the same, which is known as self-attention [13]. The number of vanilla Transformers in feature extraction module is 5 and the number of KS-Transformers in deep fusion module is 2. Eight attention heads are used in multi-head attention. The number of CCABs used in modality interaction module will be discussed later.

3.3. Experimental results and analysis

3.3.1. Key-sparse attention analysis

To demonstrate the effectiveness of the key-sparse attention, we consider a sample in IEMOCAP and compare the attention weights in vanilla Transformer and KS-Transformer by visualization. As shown in Figure 5, the vanilla Transformer takes note of all the words, including the noisy words which are not related to emotion, and trends to over-fitting. However, the KS-Transformer makes the connections from dense to sparse, which is able to ignore most of the noises and focus more on the emotional information. Meanwhile, the sparsity in KS-Transformer can reduces the complexity in the model and alleviates over-fitting.
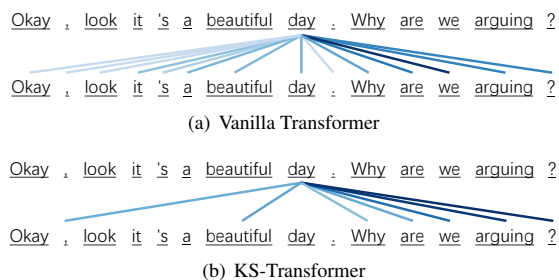
---

[1] https://github.com/pytorch/fairseq

(a) Vanilla Transformer

(b) KS-Transformer

Fig. 5. Visualization of the attention weights.

(a) IEMOCAP   (b) LSSED

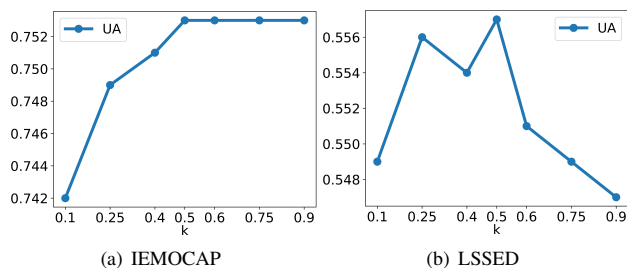**Fig. 6**. Effects of hyperparameter k.

To explore the optimal sparsity in KS-Transformer, we vary k from 0.1 to 0.9. The larger k we set, the less attention weights are reset to zero and less sparsity we have. Because LSSED suffers from sample imbalance, we use unweighted accuracy (UA) as criterion. The results are shown in Figure 6.

Since IEMOCAP is a relatively small corpus, the model is prone to over-fitting when k is larger than 0.5, causing the UA scores to remain constant. However, on the large-scale dataset LSSED, a significant drop is appeared when k is larger than 0.5 because of the redundant information. In contrast, when k is smaller than 0.5, the model uses too little information and might converge to an unsatisfactory local minimum. Considering the UA performance curves on IEMOCAP and LSSED corpora, k is set to 0.5, which means 50% attention weights are reset to zero in each KS-Transformer as default.

3.3.2. Multimodal interaction analysis

Modality interaction is vital for multimodal system. To investigate the effectiveness of the stack of CCABs, we change the number of CCABs used from zero to four, where zero means that the modality interaction module is removed, and the results are shown in Table 1. Weighted accuracy (WA) and UA are used as criteria. It should be noted that the number of CCABs used represents the times of interactions performed.

From Table 1, we show that the interaction between different modalities is shallow and insufficient when only one CCAB is applied. The performance improves as the number of CCABs increases. The best performances are obtained

when the number is three, which confirms the effectiveness of CCAB and the necessity of multiple times interactions.

Table 1. Performances of different number of CCABs in modality interaction module on IEMOCAP and LSSED.

Amount	IEMOCAP		LSSED	
	WA	**UA**	**WA**	**UA**
0	0.726	0.734	0.647	0.540
1	0.724	0.735	0.648	0.544
2	0.731	0.740	**0.651**	0.554
3	**0.743**	**0.753**	0.650	**0.557**
4	0.742	0.751	0.648	0.555

3.3.3. Comparison with some known systems

Table 2 gives the performance comparison among the proposed method with some known systems on IEMOCAP and LSSED, in which, all the systems apply audio and text as inputs except that PyResNet [20] only takes audio information.

From Table 2, it can be observed that our method gives the best WA and UA on IEMOCAP. Moreover, our method achieves the highest UA on LSSED, where UA is a more important criterion because of the sample imbalance issue.

Table 2. Comparison results on IEMOCAP and LSSED.

Dataset	Methods	Year	WA	UA
IEMOCAP	CMA [6]	2020	-	0.728
	STSER [18]	2020	0.711	0.721
	GBAN [17]	2020	0.724	0.701
	Ours	2021	**0.743**	**0.753**
LSSED	CMA	2020	0.616[#]	0.489[#]
	STSER	2020	**0.651**[#]	0.512[#]
	PyResNet	2021	0.624	0.429
	Ours	2021	0.650	**0.557**

[#] LSSED is a new released dataset. Author provides these results by reproducing the corresponding methods and training and testing them on LSSED dataset.

4. CONCLUSION

In this paper, KS-Transformer, using a novel key-sparse attention mechanism, has been proposed for speech emotion recognition. Only the emotion related speech frames in audio or words in text can be considered and assigned with attention weights. And based on KS-Transformer, we further present CCAB to fuse different modalities and achieve deep interaction. Experimental results on IEMOCAP and LSSED demonstrate the effectiveness of KS-Transformer and CCAB. In the future, we plan to combine more modalities to further improve the system performance.

6. REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[2] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Defense Science Research Conference and Expo (DSR)*, 2011, pp. 1–5.

[3] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-Modal Attention for Speech Emotion Recognition," in *Proc. Interspeech 2020*, pp. 364–368.

[4] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multimodal learning better than single (provably)," *arXiv preprint arXiv:2106.04538*, 2021.

[5] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118.

[6] D. N. Krishna and A. Patil, "Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks," in *Proc. Interspeech 2020*, pp. 4243–4247.

[7] R. Peri, S. Parthasarathy, C. Bradshaw, and S. Sundaram, "Disentanglement for audio-visual emotion recognition using multitask setup," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6344–6348.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition.," in *Proc. Interspeech 2019*, pp. 3465–3469.

[11] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176274–176285, 2020.

[12] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Proc. Interspeech 2020*, pp. 3755–3759.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-Attention for Speech Emotion Recognition," in *Proc. Interspeech 2019*, pp. 2578–2582.

[15] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3507–3511.

[16] C. Busso, M. Bulut, C.-C Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[17] P. Liu, K. Li, and H. Meng, "Group Gated Fusion on Attention-Based Bidirectional Alignment for Multimodal Emotion Recognition," in *Proc. Interspeech 2020*, pp. 379–383.

[18] M. Chen and X. Zhao, "A Multi-Scale Fusion Framework for Bimodal Speech Emotion Recognition," in *Proc. Interspeech 2020*, pp. 374–378.

[19] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 02, pp. 1359–1367, Apr. 2020.

[20] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: A large-scale dataset and benchmark for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 641–645.