HIERARCHICAL ATTENTION-BASED TEMPORAL CONVOLUTIONAL NETWORKS FOR EEG-BASED EMOTION RECOGNITION

Chao Li[1], *Boyang Chen*[1], *Ziping Zhao*[1], *Nicholas Cummins*[2], *Björn W. Schuller*[1,3,4]

[1] College of Computer and Information Engineering, Tianjin Normal University, China
[2] Department of Biostatistics and Health Informatics, IoPPN, Kings College London, UK
[3] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[4] GLAM- Group on Language, Audio & Music, Imperial College London, UK

ABSTRACT

EEG-based emotion recognition is an effective way to infer the inner emotional state of human beings. Recently, deep learning methods, particularly long short-term memory recurrent neural networks (LSTM-RNNs), have made encouraging progress for in the field of emotion recognition. However, the LSTM-RNNs are time-consuming and have difficulty avoiding the problem of exploding/vanishing gradients when during training. In addition, EEG-based emotion recognition often suffers due to the existence of silent and emotional irrelevant frames from intra-channel. Not all channels carry the same emotional discriminative information. In order to tackle these problems, a *hierarchical attention-based temporal convolutional networks* (HATCN) for efficient EEG-based emotion recognition is proposed. Firstly, a spectrogram representation is generated from raw EEG signals in each channel to capture their time and frequency information. Secondly, temporal convolutional networks (TCNs) are utilised to automatically learn more robust/intrinsic long-term dynamic characters in emotion response. Next, a hierarchical attention mechanism is investigated that aggregates the emotional information at both the frame and channel level. The experimental results on the DEAP dataset show that our method achieves an average recognition accuracy of 0.716 and an F1-score of 0.642 over four emotional dimensions and outperforms other state-of-the-art methods in a user-independent scenario.

Index Terms— emotion recognition, EEG signals, temporal convolutional networks, hierarchical attention mechanism

1. INTRODUCTION

Emotional intelligence [1] plays an important role in Human-Computer Interaction (HCI), which can provide users with a smoother interface and give them appropriate feedback or recommendation. For example, detecting the user's emotional state can be used for adaptive music recommendation, to suggest music clips that match the current emotional state or to help the user overcome negative emotions.

Due to the objectiveness of the central nervous system (C-NS) on the human emotional presentation, many recent works [2] have emerged to explore the relationships between EEG signals and their corresponding emotional states. Conventional approaches to recognising emotional states rely heavily on hand-crafted features, which requires professional domain knowledge and extensive preprocessing for the specific task [3]. Owing to the success attained by deep learning techniques in classification tasks, convolutional neural networks (CNNs) are increasingly utilised to automatically capture feature representation for EEG-based emotion recognition. Multi-channel EEG signals are usually converted into 2D images and fed into CNNs to facilitate the classification of users' emotions using EEG signals [4]. Hierarchical convolutional neural networks (HCNNs) [5] have further been proposed to classify users' emotional states; under this approach, differential entropy features from different channels are used as two-dimensional maps to train the HCNNs. Although these methods have improved the performance of EEG-based emotion recognition to a certain extent, CNN-based approaches still lack the ability to model the temporal dynamics of emotional response.

Considering that EEG signals are essentially multi-channel time-series signals, another intuitive solution for emotion recognition is to use recurrent neural networks (RNNs) to obtain long-term dependencies in emotional representation. RNN-based approaches have demonstrated their ability to capture temporal information in the EEG-based emotion recognition context [3]. Moreover, recent studies have focused on designing hybrid architectures using a com-

The work presented in this article was substantially supported by the National Natural Science Foundation of China (Grant No: 62071330, 61702370, 61902282), the National Science Fund for Distinguished Young Scholars (Grant No: 61425017), the Key Program of the National Natural Science Foundation of China (Grant No: 61831022), the Key Program of the Natural Science Foundation of Tianjin (Grant No.18JCZDJC36300), the technology plan of Tianjin (Grant No: 18ZXRHSY00100). Ziping Zhao is the corresponding author (ztianjin@126.com).

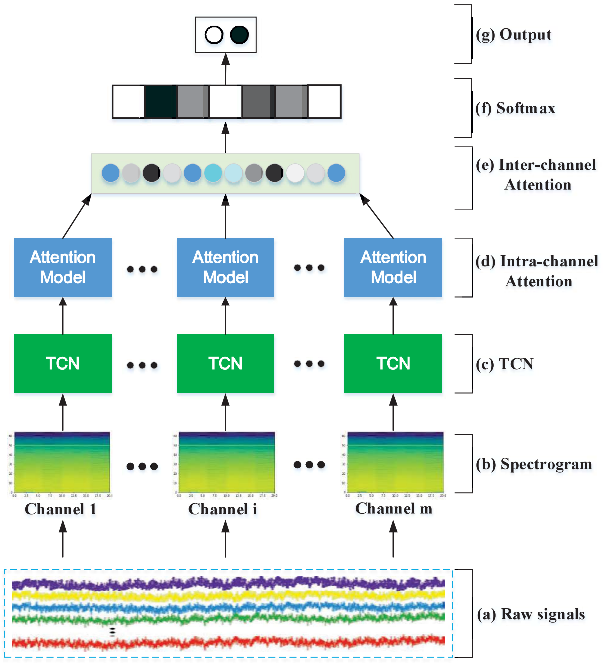

Fig. 1. The framework of the hierarchical attention-based temporal convolutional networks (HATCN) .

bination of CNN and RNN approaches, such as cascaded or parallel convolutional recurrent networks [6, 7] . However, such frameworks are affected by certain limitations. In RN-N, the predictions of the emotional label must wait until all predecessors have finished their tasks [8]. Moreover, it is difficult for LSTM-RNNs to avoid the exploding/vanishing gradient problem when input sequences are long [9]. Recently, due to their advantageous parallelism, flexible receptive field and stable gradient, temporal convolutional networks [8] have proven effective at capturing long range patterns [10].

In addition, benefiting from flexibility in the decoding phase, attention mechanisms have demonstrated their capability to extract key sequential information and automatically skip redundant information for emotion recognition tasks [11]. However, limited research has been undertaken to explore the combination of temporal convolutional network (TCN) and a hierarchical attention mechanism for extracting emotionally salient information from spectrograms for the task of EEG-based emotion recognition. Motivated by the above analysis, in this work, we propose hierarchical attention-based temporal convolutional networks (HATCN) for efficient EEG-based emotion recognition that operates by leveraging a hierarchical attention mechanism based on TCN. The main contributions of this paper are as follows:

1) Compared with CNN-/RNN-based approaches, TCN is more capable of efficiently learning spatial-temporal features from spectrogram representation and modelling the temporal dependencies between their activations;

2) To better capture salient emotional information from both an intra- and inter-channel perspective, a hierarchical attention mechanism is utilised that can allocate importance to EEG signals at the frame and channel level and aggregate this information to form a higher-level representation.

3) Experimental results indicate that our method outperforms the existing methods in a user-independent scenario.

2. METHODOLOGY

2.1. System overview

As illustrated in Figure 1, in our model, a spectrogram is first generated from the raw signal for each channel, which provides important time and frequency information. Secondly, the generated spectrogram is fed into a TCN to automatically capture the spatial-temporal feature representation. Subsequently, a hierarchical attention mechanism is designed for extracting the emotional information from both within and between the EEG signal channels. Finally, a softmax classification is utilised to predict the final emotion state.

2.2. Spectrogram representation

Spectrogram representation is generally considered as sufficiently discriminative for these purposes, as it captures the different characteristics of signals by employing distinctive patterns.Here, EEG signals in each channel are transformed into spectrogram images by means of STFT using a Hamming window function (window length: $3\,s$, overlap: $1/8\,s$).

2.3. Temporal Convolutional Network (TCN)

To better capture intrinsic time-frequency information from the spectrogram, a temporal convolutional network [8] is utilised to learn the temporal dynamics representation. A TCN cell consists of three parts, namely *causal convolutions*, *dilated convolutions* and *residual connections*. In causal convolutions, information cannot be passed from the future to the past. Moreover, given that sequence modelling should be capable of looking 'very far' into the past, dilated convolutions are employed to enable an exponentially large receptive field. More specifically, for a sequence input x $\in \mathbb{R}^n$, the dilated convolution operation F on element s of the sequence is defined as follows:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i}, \qquad (1)$$

where d denotes the dilation factor, k denotes the filter size, and $s - d \cdot i$ accounts for the direction of the past. In formula 1, using a larger dilation enables an output at the top level to represent a wider range of inputs, which effectively expands the receptive field. This ensures that there exists a filter capable of hitting every input within the effective history, while also allowing for an extremely large effective history using deep networks [8].

To facilitate the stabilisation of deeper and larger TCNs, a residual block [12] is employed in place of one convolutional layer, enabling the effective learning of modifications to the identity mapping rather than the entire transformation.

2.4. Hierarchical Attention Mechanism

Intra-channel Attention. The goal of intra-channel attention is to find the key frames that are more emotionally informative than others. The output of TCN h_{mt} from each channel is input into a full connection layer, and its hidden representation u_{mt} is calculated by formula (2). Next, the similarity between u_{mt} and the intra-channel context vector u_{intra} is obtained and normalised weight α_{mt} denoting the importance of the frame x_{mt}, which can be measured by a softmax function as in formula (3). Subsequently, we calculate the epoch vector e_m as a weighted sum of the sample annotations based on their weights via formula (4). The vector u_{intra} can be regarded as a frame-level feature representation of a fixed query, namely 'what are the most important frames'.

$$u_{mt} = \tanh(W_{\text{intra}} h_{mt} + b_{\text{intra}}), \tag{2}$$

$$\alpha_{mt} = \frac{\exp(u_{mt}^T u_{\text{intra}})}{\sum_{t=1}^{T} \exp(u_{mt}^T u_{\text{intra}})}, \tag{3}$$

$$e_m = \sum_{t=1}^{T} \alpha_{mt} h_{mt}. \tag{4}$$

Inter-channel Attention. Similar to intra-channel attention, the inter-channel attention layer is employed to reward the clue modalities that contribute to correctly classifying emotional states. Similarly, the e_m from each channel is fed into a dense layer, and its hidden representation u_m is calculated by formula (5). The similarity between u_m and the inter-channel context vector u_{inter} is then computed and the normalised importance weight α_m is measured via a softmax function as formula (6). Then, we next obtain the sequence vector v as a weighted sum of the epoch annotations based on their weights via formula (7), which fuses all the information of epochs in a sequence. The context vector u_{inter}, the default weights W_{inter}, and the bias vector are randomly initialised and fine-tuned during the training process:

$$u_m = \tanh(W_{\text{inter}} e_m + b_{\text{inter}}), \tag{5}$$

$$\alpha_m = \frac{\exp(u_m^T u_{\text{inter}})}{\sum_{m=1}^{M} \exp(u_m^T u_{\text{inter}})}, \tag{6}$$

$$v = \sum_{m=1}^{M} \alpha_m e_m. \tag{7}$$

2.5. Emotion classification

Vector v represents a learnt feature vector of the EEG sequence, which includes more discriminative and robust emotional information from both intra- and inter-channel perspectives. Finally, the emotion can be classified by means of the following equation:

$$p = soft\max(W_c v + b_c) \tag{8}$$

In short, our model is trained by minimising the cross-entropy between the predicted label and the real label.

3. EXPERIMENTS AND ANALYSIS

3.1. DEAP Dataset

In this paper, we used the DEAP dataset [13] to validate our proposed method. In this dataset, 32-channel EEG and 8-channel peripheral physiological signals from 32 subjects were recorded, when she/he watched selectively 40 one-minute music videos from 120 emotional stimuli. The emotional label for each music video is annotated by self-report on four dimensions: Arousal, Valence, Dominance, and Liking. The original affective label scales (from 1 to 9) are mapped into either a high level, or low level category by thresholding at level 5 (high level$\geq$5, low level$<$5). In this paper, we use EEG signals in this dataset to build our model.

3.2. Experimental Setup and Evaluation Metrics

In this paper, leave-one-subject-out cross-validation is used to evaluate the performance of several methods for the purposes of user-independent emotion recognition. We compare the performance of our methods with the following existing methods on the DEAP dataset: MKL [14], Bayes classifier [15], SVM [16], and CNN-, LSTM-, and TCN-based methods. All deep learning-based approaches are implemented using the Tensorflow library. For each classification algorithm, we manually tune its parameters to achieve optimal performance. Accuracy (ACC) and macro-F1 score (F1-score) are the evaluation measures employed to assess emotion recognition performance. We evaluate the recognition performance on the arousal, valence, dominance, and liking dimensions. Further details on our proposed model are given below: (a) the temporal convolutional network contains 32 nodes; (b) the dropout rate is set to 0.5; (c) batch normalisation techniques and a PReLU activation function are applied to prevent overfitting; (d) the Adam optimiser with a learning rate of 0.001 is used for training; (e) the batch size is set to 64.

3.3. Experimental Results

Table 1 presents the experimental results of existing methods and our proposed method on the DEAP dataset. We can

Table 1. Comparison results in terms of ACC and F1-score with other existing methods on the DEAP dataset.

Classifier	**Arousal**		**Valence**		**Dominance**		**Liking**		**Average**	
	ACC	**F1-score**	**ACC**	**F1-score**	**ACC**	**F1-score**	**ACC**	**F1-score**	**ACC**	**F1-score**
MKL [14]	0.580	0.520	0.590	0.550	N/A	N/A	0.660	0.510	0.610	0.527
Bayes Classifier [15]	0.570	N/A	0.620	N/A	N/A	N/A	N/A	N/A	0.595	N/A
SVM [16]	0.605	0.570	0.656	0.645	0.583	0.533	0.583	0.533	0.607	0.570
1-layer CNN+DNN	0.616	0.561	0.617	0.591	0.601	0.521	0.662	0.581	0.624	0.564
2-layer CNN+DNN	0.605	0.563	0.588	0.558	0.630	0.564	0.632	0.537	0.614	0.556
3-layer CNN+DNN	0.599	0.546	0.595	0.571	0.607	0.529	0.652	0.542	0.613	0.547
1-layer BLSTM+DNN	0.643	0.587	0.627	0.592	0.670	0.586	0.698	0.573	0.660	0.585
2-layer BLSTM+DNN	0.643	0.587	0.630	0.595	0.661	0.560	0.696	0.564	0.658	0.577
1-layer BLSTM+Attention	0.655	0.596	0.629	0.583	0.671	0.565	0.712	0.584	0.667	0.582
2-layer BLSTM+Attention	0.637	0.588	0.638	0.595	0.671	0.587	0.685	0.576	0.658	0.587
1-layer TCN+DNN	0.649	0.601	0.634	0.581	0.679	0.597	0.705	0.587	0.667	0.592
2-layer TCN+DNN	0.643	0.587	0.627	0.592	0.670	0.586	0.698	0.573	0.660	0.585
1-layer TCN+Attention	0.701	0.623	0.672	0.618	0.703	**0.623**	0.732	0.630	0.702	0.624
Ours	**0.710**	**0.646**	**0.691**	**0.657**	**0.719**	0.621	**0.742**	**0.645**	**0.716**	**0.642**

Table 2. Overview of the studies for emotion recognition on DEAP dataset

Reference	**Year**	**Features**	**Classifiers**	**Evaluation Methods**	**Accuracy**		
					Arousal	**Valence**	**Dominance**
Koelstra et al. [13]	2012	Power spectral features	NBC	leave-one-trial-out validation	0.620	0.576	N/A
Li et al. [17]	2015	DBN features	SVM	10-fold cross-validation	0.642	0.584	0.658
Zhuang et al. [18]	2017	The first difference of time series	SVM	leave-one-trial-out validation	0.691	**0.720**	N/A
Arevalillo-Herráez et al [19]	2019	Power spectral features	SVM	leave-one-subject-out validation	0.540	0.640	N/A
Hao Chao et al. [20]	2019	Multiband Feature Matrix	CapsNet	10-fold cross-validation	0.683	0.667	0.673
Ours		Spectrogram representation	HATCN	leave-one-subject-out validation	**0.710**	0.691	**0.719**

clearly conclude the following: 1) Apart from F1-score for dominance, our method outperforms all other methods in all evaluation measures for all emotional dimensions. 2) Deep learning methods achieve comparable or superior results to shallow learning methods. 3) Due to their lack of sequence modelling ability, CNN-based methods achieve the worst performance out of all deep learning methods. 4) Among the fusion strategies with DNNs, the TCN-based methods achieve better recognition results than CNN- and LSTM-based methods, which implies that TCN is better suited to capture high-level feature representation. 5) Compared with TCN+DNN methods, our proposed method achieves the best performance overall, which implies the effectiveness of hierarchical attention mechanisms for multi-channel emotion recognition.

3.4. Discussion

In general, our proposed model achieves the best performance in terms of accuracy and F1-score. The main reason for this is that the TCN is capable of extracting spatial and temporal features, while the hierarchical attention mechanism effectively retains the key emotional information from both intra- and inter-channel EEG signals. In addition, due to the limitations of hand-crafted features in terms of spatial and temporal representation, shallow learning methods find it difficult to achieve satisfactory recognition results. The results from deep learning methods imply that it is crucial to use either spatial or temporal information to boost emotion recognition. In terms of performance improvements, it is clear that the hierarchical attention mechanism improves the recognition accuracy of the TCN modules.

Moreover, in Table 2, we provide an overview of existing studies with different evaluation methods on the DEAP dataset for EEG-based emotion recognition. This comparison reveals that it is more difficult to establish a user-independent model with the LOSOCV method because of the influence of individual differencesEven under these conditions, our method still improves the accuracy for the valence and arousal dimensions compared with most existing methods.

4. CONCLUSION

In this letter, we proposed a hierarchical attention-based deep network architecture for efficient EEG-based emotion recognition, in which a temporal convolutional network was utilised to learn high-level feature representation and model temporal dependencies from spectrograms. The hierarchical attention mechanism was found to be capable of capturing discriminative emotional information from both frame- and modality-level of EEG signals, thereby improving the performance of emotion recognition systems. Experimental results demonstrated that our proposed model achieves competitive results on the DEAP dataset. In future efforts, meta-learning strategies shall also be investigated to further boost the performance of the emotion recognition system.

5. REFERENCES

[1] R.W Picard, *Affective Computing*, Cambrige, MA:MIT Press, 1997.

[2] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, June 2017.

[3] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu, "Emotion recognition using multimodal residual lstm network," in *Proc. 27th ACM International Conference on Multimedia (ACM MM)*, Nice, France, 2019, pp. 176–183.

[4] Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshu Mittal, and Samit Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on deap dataset," in *Proc. 29th AAAI Conference on Innovative Applications (IAAI-17)*, San Francisco, California, USA, 2017, p. 47464752.

[5] Jinpeng Li, Zhaoxiang Zhang, and Huiguang He, "Hierarchical convolutional neural networks for eeg-based emotion recognition," *Cognitive Computation*, vol. 10, no. 2, pp. 368–380, Dec 2018.

[6] Xiang Li, Dawei Song, Peng Zhang, Guangliang Yu, Yuexian Hou, and Bin Hu, "Emotion recognition from multi-channel eeg data through convolutional recurrent neural network," in *Proc. 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Shenzhen, China, 2016, pp. 352–359.

[7] Yilong Yang, Qingfeng Wu, Ming Qiu, Yingdong Wang, and Xiaowei Chen, "Emotion recognition from multichannel eeg through parallel convolutional recurrent neural network," in *Proc. 2018 International Joint Conference on Neural Networks (IJCNN)*, Rio, Brazil, 2018, pp. 1–7.

[8] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[9] James Martens and Ilya Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *Proc. International Conference on Machine Learning (ICML)*, Bellevue, Washington, USA, 2011, pp. 1033–1040.

[10] Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, "Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[11] Chao Li, Zhongtian Bao, Linhao Li, and Ziping Zhao, "Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multimodal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, pp. 102185, May 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[13] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, June 2012.

[14] Melih Kandemir, Akos Vetek, Mehmet Gönen, Arto Klami, and Samuel Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, Sep 2014.

[15] C Godin, F Prost-Boucle, A Campagne, S Charbonnier, S Bonnet, and A Vidal, "Selection of the most relevant physiological features for classifying emotion," in *Proc. 2nd International Conference on Physiological Computing Systems (PhyCS 2015)*, Angers, France, 2015, pp. 17–25.

[16] Mohammad Soleymani, Frank Villaro-Dixon, Thierry Pun, and Guillaume Chanel, "Toolbox for emotional feature extraction from physiological signals (teap)," *Frontiers in ICT*, vol. 4, pp. 1, Feb 2017.

[17] Xiang Li, Peng Zhang, Dawei Song, Guangliang Yu, Yuexian Hou, and Bin Hu, "Eeg based emotion identification using unsupervised deep feature learning," in *Proc. SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research*, Santiago, Chile, 2015.

[18] Ning Zhuang, Ying Zeng, Li Tong, Chi Zhang, Hanming Zhang, and Bin Yan, "Emotion recognition from eeg signals using multidimensional information in emd domain," *BioMed Research International*, vol. 2017, Aug 2017.

[19] Miguel Arevalilloherraez, Maximo Cobos, Sandra Roger, and Miguel Garciapineda, "Combining intersubject modeling with a subject-based data transformation to improve affect recognition from eeg signals.," *Sensors*, vol. 19, no. 13, pp. 2999, 2019.

[20] Hao Chao, Liang Dong, Yongli Liu, and Baoyun Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, pp. 2212, May 2019.