Neurocomputing 463 (2021) 177–184

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

ScalingNet: Extracting features from raw EEG data for emotion recognition

Jingzhao Hu [a], Chen Wang [a], Qiaomei Jia [a], Qirong Bu [a], Richard Sutcliffe [a,b,*], Jun Feng [a,*]

[a] School of Information Science and Technology, Northwest University, Xian 710127, China
[b] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

ARTICLE INFO

Article history:
Received 22 January 2021
Revised 21 June 2021
Accepted 8 August 2021
Available online 11 August 2021
Communicated by Zidong Wang

Keywords:
Deep learning
Convolutional Neural Networks
EEG
Emotion recognition
ScalingNet

ABSTRACT

Convolutional Neural Networks (CNNs) have achieved remarkable performance breakthroughs in a variety of tasks. Recently, CNN-based methods that are fed with hand-extracted EEG features have steadily improved their performance on the emotion recognition task. In this paper, we propose a novel convolutional layer, called the Scaling Layer, which can adaptively extract effective data-driven spectrogram-like features from raw EEG signals. Furthermore, it exploits convolutional kernels scaled from one data-driven pattern to exposed a frequency-like dimension to address the shortcomings of prior methods requiring hand-extracted features or their approximations. ScalingNet, the proposed neural network architecture based on the Scaling Layer, has achieved state-of-the-art results across the established DEAP and AMIGOS benchmark datasets.

1. Introduction

Emotion recognition plays a very important role in human–computer interaction [1]. By recognizing human emotions more accurately and quickly, we can interact with computers more efficiently, thus improving the quality of life [2]. Generally, expressive modalities can be used to judge human emotions, such as facial expressions, audio-visual expressions, and body language [3]. On the other hand, it has been suggested [4] that a distinction should be made between actual emotions and core affect. In order to take this into account, we can define what is being measured as a variable which is dependent on subjective scores, such as Arousal, Valence and Dominance. In recent years, more and more studies that recognize human emotions have used physiological electrical signals [5]6, such as Galvanic Skin Response (GSR), Skin Temperature (ST), ElectroCardioGram (ECG), ElectroMyogGraphy (EMG) and ElectroEncephaloGraphy (EEG). Relatively, EEG signals have the advantage that they are not usually easy to disguise or affected by medicines [7]; on the other hand, while their low Signal-to-Noise Ratio (SNR) can put high demands on an analysis algorithm, our approach has proven to be quite robust in this regard. In this work, therefore, we use EEG signals to recognize human emotions.

It has been proved that there are close correlations between human emotions and brain states [8]9. With the progress in EEG hardware equipment, it is nowadays feasible to collect EEG signals with higher and higher sampling rates [10]. Meanwhile, the processing and analysis methods of EEG signals are being explored and researched constantly [11]. In EEG-based emotion recognition, researchers mainly focus on three technical aspects. Firstly, the most widespread methods are based on feature engineering and machine learning algorithms to recognize human emotions [12]. This requires hand-extracted emotion-related features from EEG signals, such as Power Spectral Density (PSD), Differential Entropy (DE), etc. Secondly, with the development of deep learning, some methods combine feature engineering and deep neural networks, replacing machine learning classifiers with neural networks such as Convolutional Neural Networks (CNNs) [13]. Thirdly, some researchers extract data-driven features from EEG signals, and employ parameterizable data representation methods or neural networks as feature extractors [14]. While the feature extraction methods mentioned above have achieved remarkable performance on EEG based emotion recognition, there is still potential for improvement. Hand-extracted features are mostly task-related, and can require strong hypotheses and mathematically-driven theoretical support. In practice, we believe that extracting features by hand is not easy and potentially not robust.

* Corresponding author at: School of Information Science and Technology, Northwest University, Xian 710127, China.
E-mail addresses: rsutcl@nwu.edu.cn (R. Sutcliffe), fengjun@nwu.edu.cn (J. Feng).

https://doi.org/10.1016/j.neucom.2021.08.018
0925-2312/

Inspired by the shortcomings of methods using features extracted by hand, we introduce an end-to-end artificial neural network method called ScalingNet which can perform emotion recognition based only on raw EEG data, and which thus does not require such features. Instead, the Scaling Layer within ScalingNet extracts features from the signal automatically: The idea is to dynamically generate a series of convolution kernels scaled from one data-driven pattern to produce a data-driven spectrogram-like feature map from raw EEG signals. The architecture we introduce has several interesting properties: (1) It automatically extracts robust feature maps from raw EEG signals without any hand-interaction. (2) It handles any length of EEG signal without requiring data alignment. (3) It is fully convolutional. (4) It is compatible with existing neural networks, providing robust feature extraction for different downstream tasks. We validate the proposed approach on the challenging DEAP and AMIGOS benchmark datasets, achieving state-of-the-art results that highlight the potential of models for data-driven feature extraction from raw EEG signals.

2. Related work

In EEG-based emotion recognition, machine learning methods provided with hand-extracted EEG features are possibly the most widely used framework. With the development of deep learning, researchers have gradually replaced machine learning methods with deep neural networks, especially CNNs [15]. The hand-extracted EEG features are mainly time domain, frequency domain, time–frequency domain and spatial domain. The classification methods mainly include Random Forest (RF), Support Vector Machines (SVM), CNNs, Long Short-Term Memory networks (LSTMs) etc. Zheng et al. [16] extracted the time domain and frequency domain features from EEG signals, such as Differential Entropy (DE), Power Spectral Density (PSD), etc., and used SVMs for emotion classification. Liu et al. [17] extracted time domain, frequency domain and time–frequency domain features, such as Hjorth, PSD, Discrete Cosine Transform (DCT), etc., and then used k-Nearest Neighbor (KNN) and RF as classifiers. Li et al. [18] proposed to perform Continuous Wavelet Transform (CWT) on the EEG signal of each channel, convert it to scalograms, then input the construction frame into CNNs and LSTM for emotion recognition. Kim et al. [19] extracted brain asymmetry features and heart rate features, and then used ConvLSTM (a combination of CNN and LSTM) for classification.

Inspired by the powerful feature transforming abilities of neural networks, some researchers propose end-to-end frameworks for EEG based emotion recognition. Jiang et al. [20] mention that automatic feature extraction does not require a large amount of prior knowledge and yields better task-relevant representations compared to hand-extracted features. Wang et al. [21] propose an EmotionNet network for EEG-based emotion classification. It can take EEG as input and uses 3-D convolution to extract spatial and temporal features for emotion recognition. However, for general purpose network layers, it is hard to learn and extract robust features from signals. In the long run, this research field still has great potential for development. We consider that there is a need for a special neural network layer that can perform robust feature extraction from raw EEG signals. In the next section we propose such a layer, together with an associated network architecture.

3. Methodology

In this section, we will firstly present the Scaling Layer, which is a building block used to adaptively extract effective data-driven spectrogram-like features from raw EEG signals. Then we will introduce a fully Convolutional Neural Network based on the Scaling Layer. We call this network ScalingNet because its core feature is the application of the Scaling Layer.

3.1. Scaling layer

The motivation is to dynamically generate a series of convolutional kernels by scaling one data-driven pattern to different periods in order to expose a frequency-like dimension from signals. This brings the possibility of automatic adaptive extraction of effective and robust data-driven spectrogram-like features from raw EEG signals, for use in downstream tasks.

We consider a multi-kernel convolutional layer that takes a one-dimensional signal with shape $(sampling\ points, 1)$ as input and produces as output a two-dimensional spectrogram-like feature map with shape $(sampling\ points, scaling\ levels)$ by means of the following layer-wise propagation rule:

$$H^{output}(l) = \delta\Big(bias(l) + downSample(weight, l) \otimes H^{input}\Big) \tag{1}$$

where H^{input} is the input vector with shape $(time\ steps, 1)$, i.e. the one-dimensional signal. H^{output} is the matrix of activations with shape $(time\ steps, scaling\ levels)$, i.e. the data-driven spectrogram-like feature map. *bias* is the biases for the multi-kernel generated by scaling a basic kernel. $\delta(\cdot)$ denotes an activation function; *weight* is the basic kernel from which other kernels are scaled. l is a hyper-parameter that controls the scaling level.

$\otimes$ is a valid cross-correlation operator, normally defined as:

$$(f \otimes g)[n] \triangleq \sum_{m=0}^{N-1} \overline{f[m]}\, g\left[(m+n)_{\bmod N}\right] \tag{2}$$

where f is $downSample(weight, l)$, g is H^{input}.

Returning to Eq. (1), *downSample* is a pooling operator that downsamples the *weight* by an average filter with a window of size 2, doing this l times. This scales the data-driven pattern *weight* to a specific period in order to capture specific frequency-like representations from H^{input}. To ensure that the length of the downsampled *weight* is always odd, the *downSample* uses a padding of size 1 for the filter when the length of the directly downsampled *weight* is potentially even.

Furthermore, $bias(l)$ is the bias for the kernel generated at the l^{th} scaling level. $H^{output}(l)$ is the activation of l^{th} scaling level. $downSample(weight, l)$ denotes the generated kernel scaled from *weight* at l^{th} level, which recursively filters the *weight* l-times.

The steps involved in using Eqs. (1) and (2) are as follows. Assume we wish to extract features for signal H^{input} at the l^{th} scaling level. We first generate the l^{th} scaling level kernel scaled from *weight* by $downSample(weight, l)$. Then, we perform the cross-correlation operator of the scaled kernel and H^{input} by Eq. (2). Then, we add the previous result and the $bias(l)$, and then feed the sum to the activation function $\delta(\cdot)$, i.e. Eq. (1).

We repeat the above process total scaling level *tsl* times with different setups of hyper-parameter l on a range of 0 to maximum scaling level *msl*, where the maximum scaling level *msl* is the l^{th} level that makes the length of vector $downSample(weight, l)$ equal to 1, and the total scaling level $tsl = msl + 1$. Finally, we stack all extracted feature vectors into a 2D tensor to obtain the data-driven spectrogram-like feature map. In particular, in order to ensure the alignment of extracted feature vectors, the length of the basic kernel *weight* must be odd and the input signal H^{input} must be padded with $(scaledKernelLength - 1)/2$. For the backpropagation, the trainable parameters are the basic kernel *weight* and biases *bias*, which will be handled by an autograd mechanism.

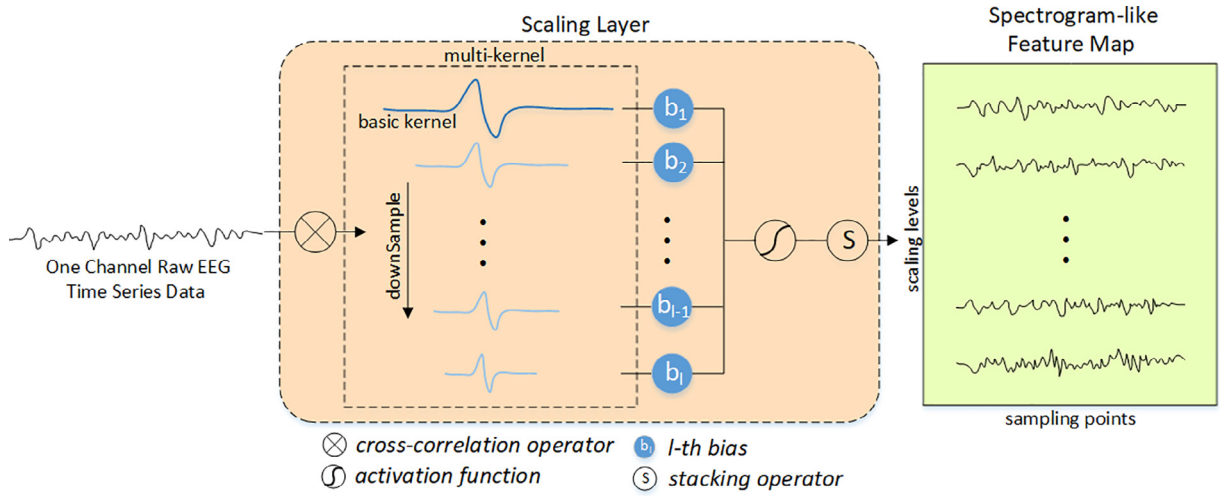


Fig. 1. The core principle of the Scaling Layer. This layer directly extracts data-driven spectrogram-like feature maps from raw EEG signals for downstream tasks. It extracts features by means of a multi-kernel generated from scaling a data-driven pattern.

The core principle of the Scaling Layer is illustrated in Fig. 1.

3.2. ScalingNet

In this subsection, we introduce ScalingNet, a neural network architecture mainly constructed by a series of parallel Scaling Layers to perform raw EEG data based emotion recognition.

The ScalingNet architecture is illustrated in Fig. 2. Considering that the Scaling Layers that are used to construct the ScalingNet extract data-driven spectrogram-like feature maps for EEG channels separately, we especially illustrate the EEG channels by stacking the data-driven spectrogram-like feature maps extracted by the Scaling Layer from EEG signals of different channels into a 3D tensor.

The EEG signals of different channels are first fed to Scaling Layers separately in order to extract data-driven spectrogram-like feature maps. Then, the feature maps extracted by the Scaling Layers are stacked into a 3D tensor along the EEG channel dimension. Next, the 3D tensor is fed into several convolutional layers to perform feature map transformation. Finally, the transformed feature maps are fed into an average global pooling layer and a linear layer to perform emotion classification. Worthily, the ScalingNet architecture robustly performs raw EEG data-based emotion recognition without requiring any hand-extracted features.

4. Experiments & results

We evaluate the performance of the proposed ScalingNet architecture on the emotion recognition task on EEG input data, using the challenging DEAP [22] and AMIGOS [23] datasets. We compare ScalingNet with previous state-of-the-art methods. We first introduce the DEAP and AMIGOS datasets, then proceed to a detailed description of the experimental setups, and finally report the experimental results.

4.1. Datasets

DEAP [24] is a challenging benchmark dataset for EEG based emotion recognition. The dataset contains EEG and physiological signals collected from 32 subjects stimulated by watching music videos. After they watch each video, the subjects immediately self-evaluate their Valence, Arousal, Dominance, and Liking, on a scale of 1–9. Each subject is asked to watch 40 videos, and 63 s of signals are collected for each video. In the dataset, the signals are downsampled by default to 128 Hz and filtered with a 4.0 Hz

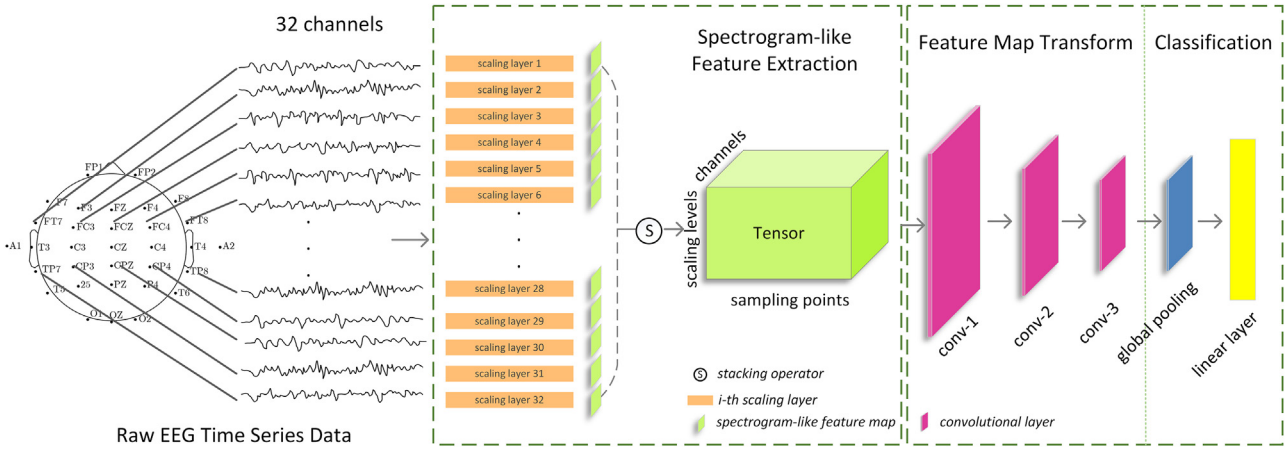


Fig. 2. The ScalingNet architecture. This is constructed by a series of parallel Scaling Layers that are followed by compact convolutional and linear layers. With the help of data-driven spectrogram-like feature maps extracted by Scaling Layers, it performs emotion recognition using raw EEG data, without any hand-extracted features.

Table 1
The hyper-parameters of the proposed ScalingNet architecture, tuned on the DEAP dataset.

Hyper-parameters	Value
batch size	32
length of *weight*	33
kernel size	3×5
number of filters	16, 8, 6
activation function	relu
loss	cross entropy
optimizer	adam

to 45.0 Hz bandpass filter. In this paper, only EEG signals are used to classify the Valence, Arousal, and Dominance by the rating threshold of 5, which closely follows the setting of [25]. Specifically, 1280 EEG samples from 32 subjects are used for three binary classification tasks of cross-subject emotion recognition.

AMIGOS [23] is another well-known dataset that can be used for EEG based emotion recognition. The dataset contains EEG signals, physiological signals, and depth videos collected from 40 subjects stimulated by watching emotional videos. After they watch each video, they immediately self-evaluate their affective levels according to a scale of 1–9, and their Valence and Arousal levels are externally rated on a scale from −1 to 1 by three annotators through the recorded face videos every 20 s. Each subject is asked to watch 20 videos, and the length of the signals depends on the length of the videos. All types of signals are default downsampled to 128 Hz and high-pass filtered with a 2.0 Hz cut-off frequency. As above, in this paper, only EEG signals are used to classify the Valence and Arousal by the rating threshold of 0, which closely follows the setting of [23]. Specifically, 12580 EEG samples from 40 subjects are used for two binary classification tasks of cross-subject emotion recognition.

4.2. Experimental setup

Fivefold, ten-fold and leave-one-subject-out (LOO) cross-validation strategies are used in the experiments. The reason is to allow direct comparison with previous state-of-the-art methods, each of which uses one of these three strategies. We manually optimize the hyper-parameters of the proposed ScalingNet architecture on the DEAP dataset, and the resulting values are shown in Table 1. In the table, 'length of *weight*' is the size of the basic kernel *weight* of the Scaling Layer in Eq. (1); 'kernel size' is the size of convolutional kernels used in the feature map transformation convolutional layers of ScalingNet, as illustrated in Fig. 2; 'number of filters' is the number of filters used in those layers (Fig. 2).

It needs to be stated that 'raw EEG' in the context of this work means that the algorithm must extract information directly from the signal itself without any human intervention. However, essential task-independent pre-processing such as epoch extraction and re-sampling is allowed.

All experiments in this paper were conducted using a GeForce RTX 2080 Ti. The machine learning framework used in this paper is PyTorch [26].

4.3. Results

The experimental results of the proposed ScalingNet architecture compared with previous state-of-the-art methods using the DEAP and AMIGOS datasets, and adopting the same evaluation strategy throughout, are shown in Tables 2 and 3. Although some researchers have investigated three dimensions, namely Arousal, Valence and Dominance, there is no validation that correlates all three with the neurophysiological responses predicted from the field of neuropsychology. In the seven comparison methods in Table 2, four of them just predict Arousal and Valence, while three

Table 2
Experimental results compared with those of previous state-of-the-art methods on the DEAP dataset. Previous approaches use various cross-validation methods, shown in the first column (Pandye et al. do not state the method used in their paper). The results for the proposed method are calculated using all three methods (5-fold, 10-fold, LOO) to allow direct comparison.

Studies	Features	Classifiers	Accuracy		
			Arousal	Valence	Dominance
Koelstra et al. (LOO)	PSD	Naive Bayes	0.6200	0.5760	–
Li et al. (10-fold)	DBN	SVM	0.6420	0.5840	0.6580
Gupta et al. (LOO)	graph	RVM	0.6700	0.6900	–
Pandye et al. (?)	VMD	DNN	0.6125	0.6250	–
Chen et al. (10-fold)	–	H-ATT-BGRU	0.6650	0.6790	–
Chao et al. (10-fold)	MFM	CapsNet	0.6828	0.6673	0.6725
Li et al. (LOO)	STFT	HATCN	0.7100	0.6901	0.7190
Ours (5-fold)	–	ScalingNet	0.6999	**0.7113**	0.7078
Ours (10-fold)	–	ScalingNet	**0.7180**	**0.7188**	**0.7367**
Ours (LOO)	–	ScalingNet	**0.7165**	**0.7132**	**0.7289**

The bold text in the table means that these our experimental results are better than the results of previous studies.

Table 3
Experiment results compared with those of previous state-of-the-art methods on the AMIGOS dataset (Luz et al. do not state their cross-validation method in the paper).

Studies	Features	Classifiers	Accuracy	
			Arousal	Valence
Juan et al. (LOO)	PSD	Naive Bayes	0.6640	0.6910
Luz et al. (?)	–	CNN	0.7350	0.6700
Yang et al. (10-fold)	VAE	SVM	0.6700	0.6880
Chao et al. (LOO)	STFT	ABLSTM	0.7280	0.6780
Ours (5-fold)	–	ScalingNet	**0.7377**	0.6880
Ours (10-fold)	–	ScalingNet	**0.7406**	**0.6952**
Ours (LOO)	–	ScalingNet	**0.7389**	**0.6928**

The bold text in the table means that these our experimental results are better than the results of previous studies.

others predict Arousal, Valence and Dominance. In this work, therefore, we report all three, to allow direct comparison with the previous results.

In Table 2, Koelstra et al. [24], proposers of the DEAP dataset, used PSD features and a Naive Bayes classifier for emotion recognition. Li et al. [27] used SVM for classifying emotions by using DBN as a feature extractor. Gupta et al. [28] used graph-theoretic features and RVM for classification. Pandye et al. [29] fed VMD features to a Deep Neural Network (DNN) for emotion classification. Chen et al. [30] proposed H-AAT-BGRU to classify emotions. Chao et al. [31] extracted MFM features and used CapsNet as a classifier for emotion recognition. Li et al. [32] fed spectrogram representations to HATCN for emotion recognition.

In Table 3, Juan et al. [23], proposers of the AMIGOS dataset, employed PSD features and a Gaussian naive Bayes classifier for emotion recognition. Luz et al. [33] used a CNN followed by a DNN to classify emotions. Yang et al. [34] used SVM for classifying emotions by using VAE as a feature extractor. Chao et al. [35] fed spectrograms to the attention-based bidirectional LSTM-RNN they proposed for emotion classification.

The results in Tables 2 and 3 show that the 5-fold/10-fold/LOO accuracies of the proposed method in this paper are 69.99%/71.80%/71.65%, 71.13%/71.88%/71.32%, 70.78%/73.67%/72.89% for Arousal, Valence, Dominance on the DEAP dataset, and 73.77%/74.06%/73.89%, 0.6880%/69.52%/ 69.28% for Arousal, Valence on the AMIGOS dataset, respectively. Using the matching cross-validation figure, these are all higher than the previous state-of-the-art studies. This indicates that the proposed ScalingNet architecture is effective and feasible for EEG data based emotion recognition.

The results above demonstrate that the spectrogram-like feature maps extracted by the Scaling Layers in ScalingNet can efficiently represent task-related information from the raw EEG signals. Compared to the hand-extracted features and general purpose network layers, in addition to not requiring any prior knowledge, the data-driven spectrogram-like features extracted by the Scaling Layer through its multiple kernels, scaled from the learned task-related patterns, can contain better representations dedicated to downstream tasks. A more detailed exploration of the Scaling Layer and ScalingNet will be presented in the next section.

5. Discussion

In this section, we have designed a series of experiments to explore the properties of the Scaling Layer and ScalingNet, to verify its contribution through ablation experiments, and to visualize the data-driven spectrogram-like feature maps extracted by the Scaling Layers.

Since the Scaling Layer handles any length of EEG signal without requiring data alignment, we can arbitrarily adjust the length of the basic kernel *weight* to explore the relationship between the model's capacity and its representational ability. We explore the relationship through observing the emotion recognition performance of ScalingNet with different setups of Scaling Layers. In the experiments, we deliberately select several representative values for the length of the basic kernel *weight* in the Scaling Layers. The results are shown in Table 4.

We can observe in the table that the representational capacity attains its best value when setting the length of *weight* to 33. Obviously, the value 33 is related to the datasets, and here we are more interested in the experimental results shown in Table 4 itself.

In order to verify the contribution of the proposed Scaling Layer, ablation experiments were also carried out. The results are shown in Table 5. Here, we compare the Scaling Layer with two alternatives from previous approaches, wavelet analysis, and a standard convolutional layer, to explore their relative feature extraction

Table 4
The relationship of Scaling Layers between its model capacity and its representational ability under the ScalingNet architecture and DEAP dataset.

Length of *weight*	Accuracy		
	Arousal	Valence	Dominance
129	0.6659	0.6778	0.6731
65	0.6773	0.6844	0.6886
63	0.6642	0.6882	0.6902
33	**0.6999**	**0.7113**	**0.7078**
17	0.6711	0.6726	0.6995

The bold text in the table means that these our experimental results are better than the results of previous studies.

Table 5
Ablation experiments varying the feature extractor within the same backend architecture and using the DEAP dataset.

Feature extractor	Accuracy		
	Arousal	Valence	Dominance
wavelet analysis	0.6477	0.6250	0.6734
convolutional layer	0.6574	0.6641	0.6628
scaling layer	**0.6999**	**0.7113**	**0.7078**

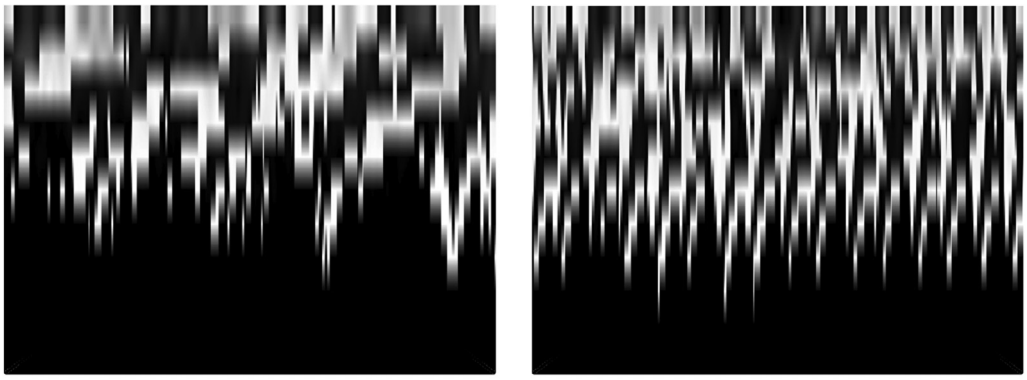
The bold text in the table means that these our experimental results are better than the results of previous studies.

capability for EEG signals. The wavelet feature extractor follows the implementation of Runia et al. [36]. We explore the capability through observing the resulting emotion recognition performance for each feature extractor. As Table 5 shows, the resulting classification accuracy under all three features, Arousal, Valence and Dominance, is best for the proposed Scaling Layer feature extractor. We can observe that the scaling layers play an important role in ScalingNet. It also indicates that the Scaling Layer extracts more robust features for EEG signals with better generalization performance.

Next, we visualize the data-driven spectrogram-like feature maps extracted by the Scaling Layers in ScalingNet, using the DEAP dataset. The feature maps are shown in Fig. 3, where the horizontal axis denotes sampling points and the vertical axis denotes the frequency-like dimension, i.e. the time and scaling levels. We can observe that Fig. 3(a) contains more low frequency-like energy and (b) contains more high frequency-like energy. It all started with one data-driven pattern which was used to generate scaled kernels in order to extract useful information. The useful learned information contained in the data-driven spectrogram-like feature maps is aggregated by the following layers and used for downstream tasks.

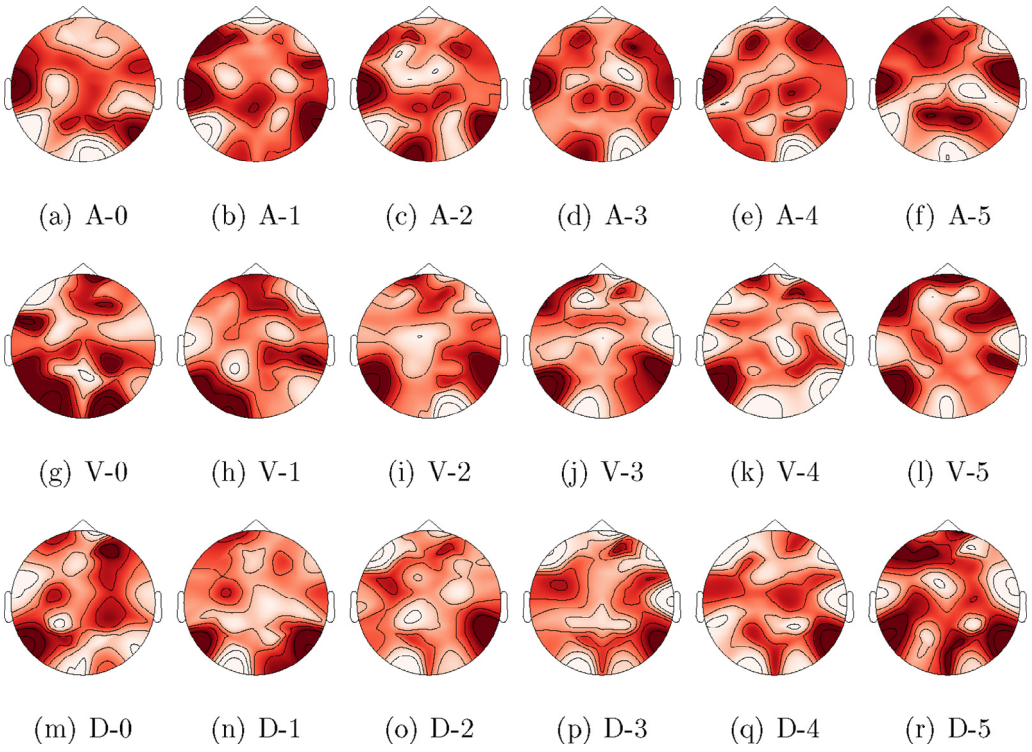
Finally, to further analyze the interpretability of the proposed Scaling Layer and ScalingNet from the perspective of brain science, we visualized the scalp topographies to see the significance of difference between positive and negative emotion groups for the features extracted by the Scaling Layers under the ScalingNet architecture. The DEAP dataset is once again used, and the results are shown in Fig. 4. The scalp topography is visualized by the $1 - p$ values calculated by the t-test method between positive and negative groups in Arousal, Valence, and Dominance across the channels and scaling levels. Here, A-0 denotes the scalp topography of Arousal at scaling level 1, D-5 stands for the scalp topography of Dominance at scaling level 6, etc. In addition, scaling levels from 0 to 5 represent a range from low frequency-like energy to high frequency-like energy.

From Fig. 4, we can observe that the brain regions used by ScalingNet to differentiate between positive and negative emotions are mainly concentrated in the prefrontal, temporal and occipital lobes.



(a) lower frequency-like energy (b) higher frequency-like energy

Fig. 3. Data-driven spectrogram-like feature maps extracted by ScalingNet Scaling Layers using the DEAP dataset.



(a) A-0 (b) A-1 (c) A-2 (d) A-3 (e) A-4 (f) A-5

(g) V-0 (h) V-1 (i) V-2 (j) V-3 (k) V-4 (l) V-5

(m) D-0 (n) D-1 (o) D-2 (p) D-3 (q) D-4 (r) D-5

Fig. 4. Scalp topographies based on $1 - p$ values calculated by the t-test method between positive and negative groups in Arousal, Valence, and Dominance across the channels and scaling levels, under the ScalingNet architecture and DEAP dataset. A-0 denotes the scalp topography of Arousal at scaling level 1, D-5 stands for the scalp topography of Dominance at scaling level 6, etc.

Among them, the prefrontal and temporal lobes have been proven to be related to emotion processing [37]. In contrast, the activation of the occipital lobe may be related to the case where the experimental paradigm used visual stimulation. Further, we can also observe that not exactly the same brain regions are attended to for different tasks and scaling levels. Notably, ScalingNet is a purely data-driven end-to-end emotion recognition method, and the brain regions of interest depend on the experimental paradigm, data, labeling, and machine learning task. With the rapid increase in the amount of data available for machine learning, it can output valuable indications relevant to brain science.

6. Conclusion

We have presented the Scaling Layer, a novel convolutional layer for extracting a spectrogram-like feature map from raw signals, and ScalingNet, a neural network that operates on raw EEG data for classification, leveraging dynamically generated convolutional kernels by scaling from one data-driven pattern. We have demonstrated that the proposed architecture can automatically extract robust data-driven spectrogram-like feature maps. The approach has been successfully applied to emotion recognition based on raw EEG data. Thus it addresses many shortcomings of prior methods based on hand-extracted features with strong hypotheses or their approximations. The ScalingNet model using Scaling Layers has successfully achieved state-of-the-art performance across two well-established emotion recognition benchmarks.

CRediT authorship contribution statement

Jingzhao Hu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization,

Writing - original draft, Writing - review & editing. **Chen Wang:** Data curation, Investigation, Software, Validation, Writing - original draft. **Qiaomei Jia:** Data curation, Investigation, Software, Validation, Writing - original draft. **Qirong Bu:** Project administration, Validation, Writing - review & editing. **Richard Sutcliffe:** Investigation, Validation, Writing - review & editing. **Jun Feng:** Funding acquisition, Project administration, Resources, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key Research and Development Program of China under grant 2017YFB1002504 and the National Natural Science Foundation of China (NSFC Grant No. 61772039 and No. 91646202).

References

[1] Y. Li, J. Huang, H. Zhou, N. Zhong, Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks, Applied Sciences 7 (10) (2017) 1060.
[2] W. Liu, W.-L. Zheng, B.-L. Lu, Multimodal emotion recognition using multimodal deep learning, arXiv preprint arXiv:1602.08225 (2016)..
[3] R. Adolphs, D. Tranel, H. Damasio, A. Damasio, Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala, Nature 372 (6507) (1994) 669–672.
[4] J.A. Russell, L.F. Barrett, Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant, Journal of Personality and Social Psychology 76 (5) (1999) 805.
[5] S. Issa, Q. Peng, X. You, Emotion classification using eeg brain signals and the broad learning system, IEEE Transactions on Systems, Man, and Cybernetics: Systems (2020).
[6] Z. Gao, Y. Li, Y. Yang, X. Wang, N. Dong, H.-D. Chiang, A gpso-optimized convolutional neural networks for eeg-based emotion recognition, Neurocomputing 380 (2020) 225–235.
[7] A. Konar, A. Chakraborty, Emotion recognition: a pattern analysis approach, Emotion Recognition: A Pattern Analysis Approach, 2015..
[8] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, G. Chen, A channel-fused dense convolutional network for eeg-based emotion recognition, IEEE Transactions on Cognitive and Developmental Systems (2020).
[9] H. Yang, J. Han, K. Min, A multi-column cnn model for emotion recognition from eeg signals, Sensors 19 (21) (2019) 4736.
[10] D. Fabiano, S. Canavan, Emotion recognition using fused physiological signals, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2019, pp. 42–48..
[11] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, Y. Zhang, et al., Eeg based emotion recognition by combining functional connectivity network and local activations, IEEE Transactions on Biomedical Engineering 66 (10) (2019) 2869–2881.
[12] W. Zheng, W. Liu, Y. Lu, B. Lu, A. Cichocki, Emotionmeter: A multimodal framework for recognizing human emotions, IEEE Transactions on Systems, Man, and Cybernetics 49 (3) (2019) 1110–1122.
[13] X. Xing, Z. Li, T. Xu, L. Shu, B. Hu, X. Xu, Sae+lstm: A new framework for emotion recognition from multi-channel eeg, Frontiers in Neurorobotics 13 (2019) 37.
[14] J. Chen, D. Jiang, Y. Zhang, A hierarchical bidirectional gru model with attention for eeg-based emotion classification, IEEE Access 7 (2019) 118530–118540.
[15] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, Deep learning-based electroencephalography analysis: a systematic review, Journal of Neural Engineering 16 (5) (2019) 051001.
[16] W. Zheng, J. Zhu, B. Lu, Identifying stable patterns over time for emotion recognition from eeg, arXiv: Human-Computer Interaction (2016)..
[17] J. Liu, H. Meng, A.K. Nandi, M. Li, Emotion detection from eeg recordings (2016) 1722–1727.
[18] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, B. Hu, Emotion recognition from multi-channel eeg data through convolutional recurrent neural network (2016) 352–359.
[19] B.H. Kim, S. Jo, Deep physiological affect network for the recognition of human emotions, IEEE Transactions on Affective Computing (2018), 1–1.
[20] J. Wang, M. Wang, Review of the emotional feature extraction and classification using eeg signals, Cognitive Robotics (2021).
[21] Y. Wang, Z. Huang, B. Mccane, P. Neo, Emotionet: A 3-d convolutional neural network for eeg-based emotion recognition (2018) 1–7..
[22] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (1) (2012) 18–31.
[23] J. A. Miran Da-Correa, M. K. Abadi, N. Sebe, I. Patras, Amigos: A dataset for affect, personality and mood research on individuals and groups, IEEE Transactions on Affective Computing (2017)..
[24] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE Transactions on Affective Computing 3 (1) (2011) 18–31.
[25] Y. Yang, Q. Wu, M. Qiu, Y. Wang, X. Chen, Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network, in, International Joint Conference on Neural Networks (IJCNN) 2018 (2018) 1–7.
[26] Pytorch. https://pytorch.org..
[27] P. Zhang, X. Li, Y. Hou, G. Yu, D. Song, B. Hu, Eeg based emotion identification using unsupervised deep feature learning (2015)..
[28] R. Gupta, K.U.R. Laghari, T.H. Falk, Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization, Neurocomputing 174 (JAN.22PT.B) (2016) 875–884..
[29] P. Pandey, K.R. Seeja, Subject independent emotion recognition from eeg using vmd and deep learning, Journal of King Saud University – Computer and Information Sciences (2019).
[30] J.X. Chen, D.M. Jiang, Y.N. Zhang, A hierarchical bidirectional gru model with attention for eeg-based emotion classification, IEEE Access 7 (2019) 118530–118540.
[31] H. Chao, L. Dong, Y. Liu, B. Lu, Emotion recognition from multiband eeg signals using capsnet, Sensors 19 (9) (2019) 2212.
[32] C. Li, B. Chen, Z. Zhao, N. Cummins, B.W. Schuller, Hierarchical attention-based temporal convolutional networks for eeg-based emotion recognition, in, in: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 1240–1244, https://doi.org/10.1109/ICASSP39728.2021.9413635.
[33] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, N. Arunkumar, Using deep convolutional neural network for emotion detection on a physiological signals dataset, amigos, Quality Control Transactions 7 (2019) 57–67.
[34] H. Yang, C. Lee, An attribute-invariant variational learning for emotion recognition using physiology, in: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 1184–1188.
[35] C. Li, Z. Bao, L. Li, Z. Zhao, Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition, Information Processing & Management 57 (3) (2020) 102185.
[36] T.F. Runia, C.G. Snoek, A.W. Smeulders, Repetition estimation, International Journal of Computer Vision 127 (9) (2019) 1361–1383.
[37] T. Iidaka, M. Omori, T. Murata, H. Kosaka, Y. Yonekura, T. Okada, N. Sadato, Neural interaction of the amygdala with the prefrontal and temporal cortices in the processing of facial expressions as revealed by fmri, Journal of Cognitive Neuroscience 13 (8) (2001) 1035–1047.

Jingzhao Hu received a B.S. degree in computer science and technology from Northwest University in 2016. Now he is a Ph.D. Candidate in Northwest University. His main research interests include brain-computer interface, deep learning, and artificial intelligence.

Chen Wang received the B.S. degree in communication engineering from Northwest University, in 2019. He is currently pursuing the M.S. degree with the School of Information Science and Technology, Northwest University, China. His research interests include electroencephalography (EEG) signal analysis, emotion recognition, mobile edge computing and deep learning.

Qiaomei Jia received a B.S. in Electronical Information Science and Technology from Northwest University China in 2019. She is studying for an M.S. degree at Northwest University. Her research is in the areas of Artificial Intelligence and Brain-Computer Interface.

Qirong Bu was born in Shaanxi, China, in 1977. He received a Ph.D. in School of Information Science and Technology from Northwest University, Xi'an, China, in 2013. He is an associate professor in School of Information Science and Technology from Northwest University, Xi'an, China. His research areas include signal processing, image processing, pattern recognition and machine learning. He has published more than 20 papers in international Journals and conferences.

Richard Sutcliffe received a Ph.D. from University of Essex in 1989. He is an Associate Professor at Northwest University China. His research interests are in the areas of Information Retrieval, Music Information Retrieval and Natural Language Processing.
Recent projects have included persuasive conversational agents, public sector message classification, analysis of classical music texts, and personality and translation ability.
He has reviewed for Artificial Intelligence Review, Computational Linguistics, Computers and the Humanities, Information Processing and Management, Information Retrieval Journal, Journal of Natural Language Engineering, Journal Traitement Automatique des Langues. Conferences he has reviewed for include ACL, CIKM, COLING, IJCNLP, LREC, NAACL-HLT, and SIGIR. He is the co-author of 108 articles and is co-editor of three books and ten conference proceeedings.

Jun Feng received a Ph.D. from City University of Hong Kong in 2006. She is a Professor in the School of Information Science and Technology at Northwest University. Her research areas include pattern recognition and machine learning, especially in the fields of brain-computer interface, medical imaging analysis and intelligent education.
Recent projects have included intelligent education based on AI and Brain-Human Interaction, and medical image analysis with deep learning. She has reviewed for many journals, including TSP, JIVP, MTAP, JDIM, CJC, JCAD, OPE, and INFPHY. Conferences she has reviewed for include IEEE-VR,MICCAI, SIGCSE, IWCSE, and CompEd. She is a member of IEEE and ACM, and is co-author of 132 articles and co-editor of three books.