MTDN: Learning Multiple Temporal Dynamics Representation for Emotional Valence Classification with EEG.

Chengxuan Tong, Yi Ding, Kevin Junliang Lim, Cuntai Guan*

***Abstract*— Emotion recognition from electroencephalogram (EEG) requires computational models to capture the crucial features of the emotional response to external stimulation. Spatial, spectral, and temporal information are relevant features for emotion recognition. However, learning temporal dynamics is a challenging task, and there is a lack of efficient approaches to capture such information. In this work, we present a deep learning framework called MTDN that is designed to capture spectral features with a filterbank module and to learn spatial features with a spatial convolution block. Multiple temporal dynamics are jointly learned with parallel long short-term memory (LSTM) embedding and self-attention modules. The LSTM module is used to embed the time segments, and then the self-attention is utilized to learn the temporal dynamics by intercorrelating every embedded time segment. Multiple temporal dynamics representations are then aggregated to form the final extracted features for classification. We experiment on a publicly available dataset, DEAP, to evaluate the performance of our proposed framework and compare MTDN with existing published results. The results demonstrate improvement over the current state-of-the-art methods on the valence dimension of the DEAP dataset.**

Index Terms—Deep learning, self-attention, electroencephalography, emotional valence

I. Introduction

Emotion is essential for humans as it affects our daily activities and decision-making. Emotion classification with electroencephalogram (EEG) has been receiving attention lately due to its objective nature, low acquisition cost, and time efficiency [1]. Emotion can be mapped into the valence, arousal, and dominance (VAD) dimensions. The valence dimension directly corresponds to positive and negative feelings [1].

EEG is a 2-dimensional time series signal, where one dimension is the channel dimension, and the other is the time dimension. The channel dimension reflects the EEG spatial information of the test subject during stimulation, while the time dimension reflects the continuous change of the subject's emotional state during external stimulation. The data can be filtered into multiple frequency bands such as delta (1-4 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (>30 Hz) bands. Each frequency band contains information about the emotional state of the subject [1]. For example, the gamma and beta bands are related to emotional valence [2], and alpha asymmetry in the frontal region is also known to be related to emotional response [3]. Besides the spatial and spectral characteristics of emotion, the temporal dynamics of EEG signals are continuous and change with time [4].

Several studies have explored emotion recognition with EEG-based BCI and achieved promising results [5]–[7]. Ding *et al*. [6] proposed TSception with multi-scaled temporal convolutional kernels to learn multiple frequency features and hemisphere kernels to learn asymmetry brain patterns for emotion recognition. Li *et al*. [5] organized EEG features as a 2-dimensional map and used hierarchical convolutional neural networks (HCNN) to extract discriminative features for emotion classification. Song *et al*. [7] proposed DGCNN, a graph-based neural network with a trainable adjacency matrix to learn the spatial features of the EEG signal for emotion classification. Deep learning on EEG is also widely applied to other BCI domains. Tong *et al*. [8] proposed TESANet, which utilized filterbank and spatial convolution to extract spectral-spatial features, and a self-attention module to learn temporal dynamics to predict the pleasantness of an olfactory dataset. Mane *et al*. [9] proposed FBCNet, which has a novel variance layer to learn the temporal information of the EEG signal. Convolutional neural network (CNN) approaches have been widely used in EEG studies [10], [11]. Such approaches utilize convolutional kernels along the channel and time dimensions to perform spatial and spectral filtering. Other method uses manual feature extraction, and then the extracted features are used as input to a classifier for class prediction [12]. The CNN and graph-based methods, namely DeepConvNet [10], EEGNet [11], FBCNet [9], TSception [6], HCNN [5], and DGCNN [7], do not have a dedicated module for temporal dynamics learning, and thus the continuous change of the emotional state is not effectively captured. Although TESANet [8] proposed using attention to learn the temporal dynamics, only a single representation of the temporal dynamics is captured, and a richer temporal context can be further explored.

There is a lack of efficient algorithms to capture all of the above-discussed features of the EEG signal. We propose to use a parallel temporal learning module on the segmented windows with overlap to address the temporal learning challenge mentioned above. In the proposed MTDN, spectral features are extracted with a filterbank module, where multiple bandpass filters are used to filter the EEG signal. We choose a filterbank from 4-40 Hz with a total of 9 bandpass filters, each with a 4 Hz bandwidth. Thus, our filtered signals include a wide range of meaningful frequencies, such as the alpha, beta, and gamma bands, which

Chengxuan Tong, Yi Ding, and Cuntai Guan are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. 50 Nanyang Ave, Singapore 639798.
Chengxuan Tong, Kevin Jun Liang Lim are with the Wilmar International 28 Biopolis Rd, Singapore 138568
* Cuntai Guan is the Corresponding Author. Email: ctguan@ntu.edu.sg

are related to emotion [2], [3]. Then, a spatial convolutional block is used to perform spatial feature learning, using a 1-D kernel with a height equal to the number of EEG channels. We then split our data into multiple overlapping time segments. The temporal dynamic learning block consists of a LSTM and a self-attention block. The temporal dynamics learning block first embeds the time segments and then learns the temporal dynamics by intercorrelating all the time segments. We utilize the self-attention mechanism [13] in our work. Parallel temporal learning blocks are stacked to learn multiple temporal dynamics jointly, providing a rich temporal representation. Finally, a fully connected (FC) layer is used for the class label prediction. We evaluate the proposed architecture with the Database for Emotion Analysis using Physiological signals (DEAP) [14]. We compare our results with the existing state-of-the-art (SOTA) and are able to obtain higher classification accuracy and F1 score.

Contributions of this study:

- MTDN, a deep learning framework to learn spatial, spectral and temporal dynamics for emotion EEG classification
- Evaluation of MTDN with DEAP dataset and comparison with existing SOTA methods.
- Ablation studies to interpret the importance of the three feature extraction blocks.

II. Methods

A. Spectral feature extraction

Let the input be (X, Y) where $X \in \mathbb{R}^{C \times T}$ is the raw EEG, $Y \in 0...(N_k - 1)$ is the label, and N_k refers to the number of classes. The filterbank is used to perform spectral filtering [15]. The filtered data is $X_f \in \mathbb{R}^{F \times C \times T}$. F, C, and T correspond to the dimension of the frequency band, EEG channel, and time points. We used 9 bandpass filters to perform the filterbank operation, which span across a wide range of frequencies and include features that are relevant to emotion recognition [1].

B. Spatial feature extraction

Spatial feature learning can be achieved by using convolutional kernels [9]–[11]. X_f is given to the spatial convolution block to generate output $X_s \in \mathbb{R}^{F*m \times T}$. m is the number of output feature maps for each convolutional filter. In this work, we set m to be 4. Depthwise convolution is used to learn 4 output feature maps for each frequency band. X_s is then applied with batch normalization and ELU activation.

C. Segmentation and dimension reduction

A segmentation operation is applied to the filtered data to split it into n time segments using an overlapping time window. This operation is done to arrange the data for the temporal learning module. The output of the segmentation operation is $X_{seg} \in \mathbb{R}^{n \times F*m \times T_{seg}}$, where T_{seg} is the length of each time window. Batch normalization and maxpooling operations are then applied to the segmented data, resulting in $X_{pool} \in \mathbb{R}^{n \times F*m \times T_{pool}}$, where T_{pool} is the number of time points after the maxpooling operation.

D. Temporal learning

X_{pool} from the previous operation is rearranged into the shape of $(n\ ,\ F * m * T_{pool})$. The data is represented by the extracted features for each time window. To maximize the effect of temporal learning, an embedding operation with a bi-layer LSTM is carried out on the sequential data. We use the hidden state of the LSTM as the embedding of the sequential data. The embedded data is denoted as $X_{emb} \in \mathbb{R}^{n \times h}$, where h is the size of the hidden embedding for each time window. X_{emb} is then given to the attention module to learn the temporal features. First, the embedded data is transformed into query (Q), key (K), and value (V) matrices with linear transformations. Q, K and V are used to evaluate the temporal features with scaled dot-product attention as shown in equation 2. This attention mechanism intercorrelates all temporal segments by evaluating the pairwise dot-product similarity. Weights are assigned to the more relevant time segments for emotional response, thus learning the temporal context of the EEG signal.

$$X_{attention} = Attention(Linear(LSTM(reshape(X_{pool})))), \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{2}$$

The temporal learning module is shown in equation 1, which utilizes a LSTM for embedding on the reshaped X_{pool}, followed by linear projection into Q, K, and V and then a scaled dot-product attention to calculate the attention output, $X_{attention} \in \mathbb{R}^{n \times d}$. To learn multiple temporal contextual information, we stack multiple temporal learning modules, and then aggregate the output from all temporal modules with a mean aggregation function as shown in equation 3. The final representation is $X_{temporal} \in \mathbb{R}^{n \times d}$, d is the output size from each attention module, k refers to the number of the temporal learning module. In our study we use 2 parallel temporal learning modules.

$$X_{temporal} = Mean\_aggregation(X^0_{attention}, ...X^{k-1}_{attention}). \tag{3}$$

E. Classification

We flatten $X_{temporal}$ into a vector of dimension $n * d$ for the final classification. A fully connected layer is used to compute the output score for the emotion class, and lastly, a softmax function to compute the final prediction of the class label.

MTDN's structure is shown in TABLE I, and the architecture is shown in Fig. 1.

III. Results and analysis

A. Dataset

We used the DEAP dataset [14] to evaluate the performance of MTDN and compared it with existing methods. The dataset consists of 32 subjects. Music video clips are used as

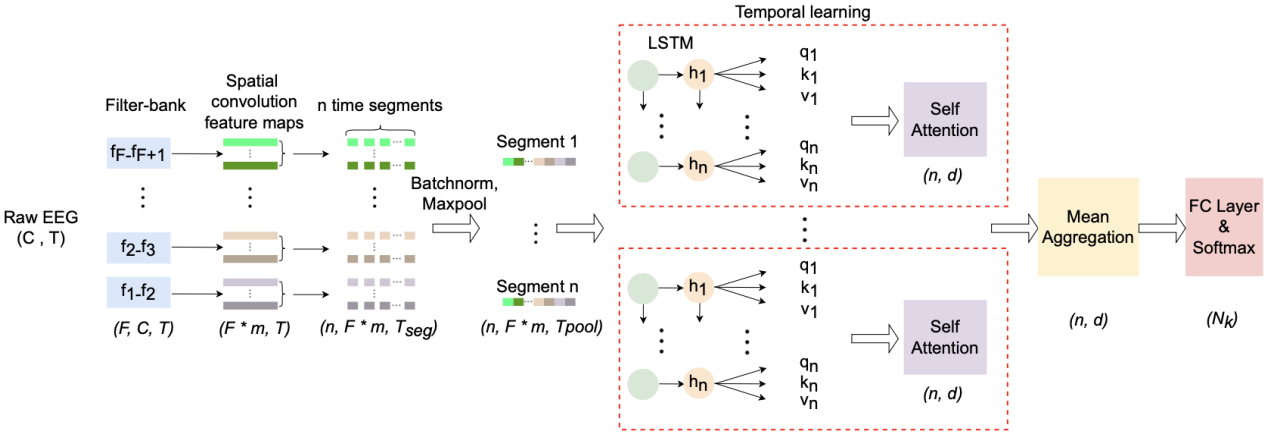

Fig. 1. MTDN structure: EEG data is fed into a filterbank and spatial convolution block for spectral and spatial feature extraction. Filter data is then split into multiple overlap time windows. Batch normalization and maxpooling are applied to the segmented data. Reshaping of the data is carried out and then given to the parallel temporal learning module. Each temporal learning module consists of LSTM embedding and scaled dot-product self-attention to find the temporal dynamics of the sequential data. The output from the parallel temporal model is then aggregated with a mean operation. Finally, a classification layer is used to predict the class label.

TABLE I
STRUCTURE OF MTDN.

Layer	Operations	Input	Output
Spectral feature extraction	Filterbank	$X(C \times T)$	$X_f(F \times C \times T)$
Spatial feature extraction	Spatial convolution block	$X_f(F \times C \times T)$	$X_s(F * m \times T)$
Segmentation and dimension reduction	Temporal segmentation	$X_s(F * m \times T)$	$X_{seg}(n \times F * m \times T_{seg})$
	Batchnorm and Maxpooling	$X_{seg}(n \times F * m \times T_{seg})$	$X_{pool}(n \times F * m \times T_{pool})$
Parallel temporal learning module	Reshape and LSTM	$X_{pool}(n \times F * m \times T_{pool})$	$X_{emb}(k \times n \times h)$
	Scaled dot-product self-attention	$X_{emb}(k \times n \times h)$	$X_{attention}(k \times n \times d)$
	Mean aggregation	$X_{attention}(k \times n \times d)$	$X_{temporal}(n \times d)$
Classification layer	Flatten, linear and softmax	$X_{temporal}(n \times d)$	$Output(N_K)$

stimuli. EEG, facial expression, and galvanic skin response are recorded. There is a total of 40 trials, where each trial lasts 1 minute with a 3 seconds pre-trial baseline. A questionnaire is given to every subject to rate emotional arousal, valence, dominance, and liking, where each dimension has 9 discrete levels. The EEG is collected with 32-channel EEG device and a sampling rate of 512 Hz.

B. Pre-processing

We follow [6] for the pre-processing pipeline. we remove the 3-seconds pre-trial baseline and downsample the data to 128 Hz. The electrooculogram is removed, and a bandpass filter of 4-45 Hz is applied. A common average reference is applied. We perform a binary classification on the valence dimension, using a score of 5 as the threshold to assign positive and negative valence to the trials. This is because the rating ranges from 1 to 9. For each trial, we split the data into non-overlapping segments of 4 seconds to obtain more data for deep learning training.

C. Experiment

We adopt the subject-dependant trial-wise 10-fold cross-validation strategy reported in [6] for this study. Every trial is split into 4 seconds of non-overlapping segments. Data-leakage problem is prevented by utilizing trial-wise 10-fold cross-validation, in which, the time segments of a trial are not simultaneously seen in both the training and test set. The setting can give us a more generalized model [6]. For each subject, 9 folds are used as the training set while 1 fold is used as the test set, then 80% of the training set is selected as training data and 20% as validation data. The model is selected based on the performance of validation data. Test data is never seen during the training phase, and final evaluation is performed on the test data.

D. Performance evaluation metrics

We use two evaluation criteria to evaluate the performance of our model, namely accuracy and F1 score. Accuracy is the ratio of correctly predicted trials over all trials. To calculate accuracy, the below expression can be used,

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{4}$$

Class label are not balanced after the processing of labels, thus F1 score is utilized as an alternate evaluation metric. The F1 score can be calculated with the following expression,

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{5}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

IV. Results and analysis

A. Experiment results

We compare MTDN with several baseline methods and the results are tabulated in Table II. MTDN achieved the highest accuracy and F1 score among all the compared models. MTDN improves by 1.17% in accuracy and 0.43% in F1 score over the second-best performing model TSception.

TABLE II
TRIAL-WISE 10-FOLD CROSS-VALIDATION RESULTS ON DEAP DATASET.

Method	Valence			
	ACC	std	F1	std
SVM [14]	55.19%	6.97%	57.87%	11.36%
KNN [12]	53.03%	9.14%	55.12%	16.27%
EEGNet [11]	54.56%	8.14%	57.61%	10.42%
ShallowConvNet [10]	59.42%	8.30%	62.26%	11.49%
DeepConvNet [10]	59.92%	7.82%	62.04%	10.23%
TESANet [8]	58.81%	9.25%	62.17%	12.26%
TSception [6]	59.14%	7.60%	62.33%	9.03%
MTDN (ours)	**60.31**%	**8.02**%	**62.76**%	**11.59**%

B. Ablation

To evaluate the contribution of the essential blocks of MTDN, we conducted an ablation study by individually removing the filterbank, spatial convolution block, and temporal learning modules, and monitoring the experiment results. The ablation results showed that the filterbank contributes the most to the classification accuracy and F1 score. A single temporal learning module also showed a drop in performance, which suggests that parallel temporal learning modules are essential for emotion recognition. All the ablation results showed a decrease in performance, which indicates that all the modules work together to give MTDN its predicting power.

TABLE III
ABLATION RESULT.

Removed operation	Valence			
	ACC	std	F1	std
w/o Filterbank	58.57%	7.87%	60.85%	12.72%
w/o Spatial convolution	59.55%	9.17%	61.96%	13.72%
w single temporal learning	58.81%	9.25%	62.17%	12.26%
MTDN (ours)	**60.31**%	**8.02**%	**62.76**%	**11.59**%

V. Conclusion

In this study, a deep learning framework for learning multiple temporal dynamics, MTDN, is proposed for emotion recognition. Parallel temporal learning blocks are utilized to jointly learn a rich temporal information representation. An experiment on the DEAP dataset is carried out to evaluate the performance. The results demonstrate MTDN outperforms the previously reported results on the DEAP dataset and improves over the compared models. An ablation study is conducted to interpret the contribution of the crucial feature extraction layers of MTDN.

Acknowledgment

I would like to thank Prof Nam-Hai Chua for his advise on the paper.

This work was supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G8b0102). This work was supported by EDB-Wilmar IPP Program.

References

[1] S. M. Alarcão and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2019.

[2] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, "Emotion classification using minimal EEG channels and frequency bands," in *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2013, pp. 21–24.

[3] J. J. B. Allen, P. M. Keune, M. Schönenberg, and R. Nusslock, "Frontal EEG alpha asymmetry and emotion: From neural underpinnings and methodological considerations to psychopathology and social cognition," *Psychophysiology*, vol. 55, no. 1, p. e13028, 2018.

[4] P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Fossati, and I. V. Mechelen, "Determinants of Emotion Duration and Underlying Psychological and Neural Mechanisms," *Emotion Review*, vol. 7, no. 4, pp. 330–335, 2015.

[5] J. Li, Z. Zhang, and H. He, "Hierarchical Convolutional Neural Networks for EEG-Based Emotion Recognition," *Cognitive Computation*, vol. 10, 04 2018.

[6] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "TSception: Capturing Temporal Dynamics and Spatial Asymmetry from EEG for Emotion Recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[7] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG Emotion Recognition Using Dynamical Graph Convolutional Neural Networks," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 532–541, 2020.

[8] C. Tong, Y. Ding, K. L. Jun Liang, Z. Zhang, H. Zhang, and C. Guan, "TESANet: Self-attention network for olfactory EEG classification," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–7.

[9] R. Mane, E. Chew, K. S. G. Chua, K. K. Ang, N. Robinson, A. P. Vinod, S. Lee, and C. Guan, "FBCNet: A Multi-view Convolutional Neural Network for Brain-Computer Interface," *CoRR*, vol. abs/2104.01233, 2021.

[10] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[11] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul 2018.

[12] M. S. Mahmud, F. Ahmed, M. Yeasin, C. Alain, and G. M. Bidelman, "Multivariate Models for Decoding Hearing Impairment using EEG Gamma-Band Power Spectral Density," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[15] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 2390–2397.