An Efficient Deep-Learning-Based Solution for the Recognition of Relative Changes in Mental Workload Using Wearable Sensors

Majd Saleh, Stéphane Paquelet, Pierre Castel, Marc Hoarau, Nico Pallamin, Daniel Lewkowicz

An efficient deep-learning-based solution for the recognition of relative changes in mental workload using wearable sensors

Majd SALEH, *Member, IEEE*
Artificial Intelligence Lab
Institute of Research and Technology b<>com
Rennes, France
Majd.SALEH@b-com.com

Stéphane PAQUELET
Artificial Intelligence Lab
b<>com
Rennes, France
Stephane.PAQUELET@b-com.com

Pierre CASTEL
Artificial Intelligence Lab
b<>com
Rennes, France
Pierre.CASTEL@b-com.com

Marc HOARAU
Human Factors Technologies Lab
b<>com
Brest, France
Mfg.HOARAU@gmail.com

Nico PALLAMIN
Human Factors Technologies Lab
b<>com
Brest, France
Nico.PALLAMIN@b-com.com

Daniel LEWKOWICZ
Human Monitoring & Data Science Department
Human Design Group
Toulouse, France
Daniel.LEWKOWICZ@humandesign.group

***Abstract*—In this work, a new solution for the automatic recognition of relative changes in mental workload is proposed. Wearable sensors were used to collect EEG, EDA, PPG and eye-tracking data from 26 human subjects while performing the n-back task with three difficulty levels $n \in \{1, 2, 3\}$. The objective is to recognize whether the mental workload is increasing, decreasing or stable by comparing the current signals' window with a previous one. The proposed 3-class classifier uses mainly CNN layers with a novel merging layer that systematically captures the interactions between local segments of the two inspected windows. In fact, it is inspired by the competitive success of both transformer- and CNN-based networks in time series classification. While the proposed solution exploits the efficiency of CNN networks, it also enjoys, similar to transformers, the capacity of capturing the interactions between local events of the sequence thanks to the proposed merging layer. In terms of accuracy, experimental results show the superiority of the proposed solution over classical CNN, BiLSTM and transformer networks on eye-direction, PPG and EEG data while its performance is comparable with the transformer networks on eye-pupil-diameter and EDA data. The average training time per epoch is considerably smaller than the ones of transformer and BiLSTM networks as shown in the experimental results.**

***Index Terms*—Mental workload (MWL), deep neural networks (DNNs), time series classification (TSC), eye-tracking, photoplethysmogram (PPG), electroencephalogram (EEG), electrodermal activity (EDA), n-back task, transformer neural network, convolutional neural network (CNN).**

I. Introduction

The concept of mental workload (MWL) describes the ability of a human operator to supply the resources required by a task [1]. Thus, given the complexity of the task and the ability of a human operator to handle it, mental workload can range from underload, normal load to overload [2]. The automatic recognition of mental workload represents a hot research topic with a wide spectrum of important applications like brain-computer interfaces, driver awareness, mental health monitoring and human safety in high workload environments [1], [2], [3].

MWL recognition is generally formulated as a classification problem [2]. Deep learning solutions for time series classification have shown extraordinary results in a wide variety of applications [17]. In fact, after the remarkable success of transformer neural networks [6] in the domain of natural language processing (NLP), transformers have been employed in time series representation and related applications. Time series transformers like the Gated Transformer Network (GTN) [9] and the Time Series Transformer (TST) [11] have achieved very good results on time series classification, including the classification of electrophysiological signals. On the other hand, some variants of CNN networks like ResNet [7], Inception [10] and Fully Convolutional Network (FCN) [12] showed a comparable performance with transformer networks. For example, FCN outperformed GTN on 5 reference datasets while GTN performed better on 4 reference datasets [9]. ResNet outperformed GTN on 5 out of 11 reference datasets. The CNN variant proposed by Franceschi et al. [16] has been compared with TST on the ECG heartbeat dataset which consists of segmented and preprocessed ECG signals for heartbeat classification. It showed an accuracy of 0.756%, just 2% less than the accuracy achieved by the TST transformer. Finally, the Inception network outperformed TST on the reference dataset IEEEPPG [13] where the objective is to estimate heart rate using PPG sensors [11].

Given these competitive results of CNNs and transformers, the motivation underlying this work is to propose a solution for the automatic recognition of MWL relative changes exploiting the efficiency advantages of the CNNs combined with the capacity of capturing the interactions between local segments

of the time series inspired by the self attention mechanism of transformers.

II. Data acquisition

Data were collected from 26 healthy participants using eye-tracking, EDA, PPG and EEG sensors. Eye tracking data were collected using *Vive Pro Eye HMD* with sampling frequency $f_s = 120$ Hz. In this study, we considered eye pupil diameter and eye direction data of the left and right eyes. EEG, PPG and EDA data were collected using the *BITalino (r)evolution Plugged Kit BT* with sampling frequency $f_s = 200$ Hz. The experimental protocol followed the n-back task [4], [3] with three difficulty levels $n \in \{1, 2, 3\}$. Participants were asked to report their subjective assessment of mental workload on classical ISA scale of 1 to 5. Raw data have been segmented into 35-second windows. Thus the dimensionality of the considered modalities are as follows:

- Eye pupil diameter data $\in \mathbb{R}^{4200\times 2}$
- Eye direction data $\in \mathbb{R}^{4200\times 6}$ for both left and right eyes each with directions on x, y and z axes.
- EEG data $\in \mathbb{R}^{7000\times 2}$ for EEG of the left and right hemispheres.
- EDA data $\in \mathbb{R}^{7000\times 1}$
- PPG data $\in \mathbb{R}^{7000\times 1}$

Then, pairs of signals have been created with the corresponding labels: increasing, stable and decreasing mental workload. Note that these pairs have been semi-randomly selected respecting two conditions: 1) left and right buffers of each pair belong to the same subject; and 2) the three classes are equally represented in the dataset (33.33% of data in each class). The resultant dataset consists in 468 samples per modality.

III. Methodology

Time series can be seen as sequences of local events. These events can be represented in real-valued vectors whose role is equivalent to token embeddings in NLP. This can be achieved using fully connected dense layers [11] or 1D convolutional layers with down-sampling (e.g. using max pooling) as discussed in [11] and recommended in [14]. We adopt the second solution in the design of the BiLSTM [8] and transformer [6] networks.

Let's illustrate the proposed architectures on the pupil diameter data as an example. The raw time series is passed through a CNN, g_{ϕ}, as illustrated in Fig. 1. Thus, the raw time series of pupil diameter data which consists in 4200 time steps and two channels is converted into a new sequence of local events with 105 time steps and 16 channels. Note that in the proposed architecture, the current and previous buffers should be processed using the same embedding layer g_{ϕ} and then the same BiLSTM/transformer h_{ψ}. The BiLSTM/transformer is not used to compress data but rather to contextualize the representation using the recurrence mechanism in the BiLSTM and the self attention mechanism in the transformer. Therefore, in order to compress the features extracted by BiLSTM/transformer before passing them to a dense classifier, a pooling layer is applied.

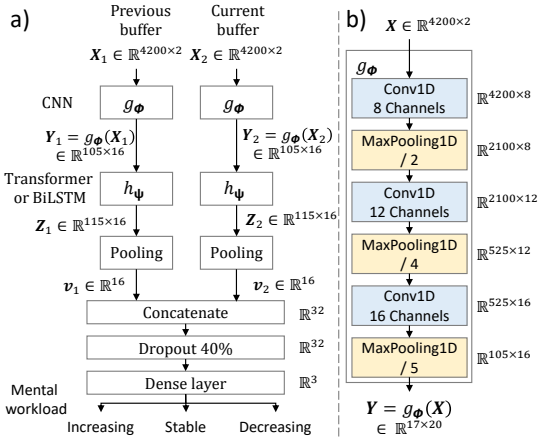

Fig. 1. The proposed architecture of the transformer- and BiLSTM-based classifiers for pupil diameter data: a) the general architecture; b) the CNN used for preparing the local events' embeddings.

Here, we describe the proposed solution, namely IG-CNN for Interaction-Grid-based CNN. The idea underlying IG-CNN is to combine the advantage of capturing the interactions between local segments of time series, inspired by transformers, and the efficiency advantage of CNNs. Fig. 2-a shows the architecture of the proposed solution. A convolutional network $f_{\boldsymbol{\theta}}$ is used to represent the the current and the previous buffers $\boldsymbol{X}_1 \in \mathbb{R}^{4200\times 2}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{4200\times 2}$, as sequences $\boldsymbol{Y}_1 \in \mathbb{R}^{17\times 20}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{17\times 20}$ of 17 time steps and 20 features. $f(\boldsymbol{\theta})$ is shown in Fig. 2-b. Theoretically, the inner product between a column of $\boldsymbol{Y}_1$ and a column of $\boldsymbol{Y}_2$ can capture the similarity between the two corresponding local segments of the time series $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in a similar manner of deducing the semantic similarity between two tokens from the inner product of (or the cosine distance between) their embeddings in NLP. This might be helpful to discriminate stable MWL from unstable one. However, in the considered application, we need instead to derive a formula that enables capturing the change direction (increasing vs. decreasing) of MWL. To this end, the interaction grid between the features extracted from the two time series is generated using the formula $\boldsymbol{Z} = (\boldsymbol{Y}_2 - \boldsymbol{Y}_1)(\boldsymbol{Y}_2 + \boldsymbol{Y}_1)^{\intercal}$ as illustrated in Fig. 2-c. The square matrix $\boldsymbol{Z} \in \mathbb{R}^{17\times 17}$ is finally flattened and passed through a fully connected layer to perform the 3-class classification.

IV. Experimental results

A. Experimental configurations

The experiments have been performed on an NVIDIA RTX A3000 GPU whose compute capability is 8.6. The RAM is 22 GB (6 GB dedicated + 16 GB shared). Keras library has been used to implement the proposed solutions with TensorFlow acting as a back-end. The dataset has been randomly split into a training set (80%) and a test set (20%). The optimizer RectifiedAdam [15] has been used with the categorical cross-entropy cost function.

The transformer network contains 4 identical blocks that

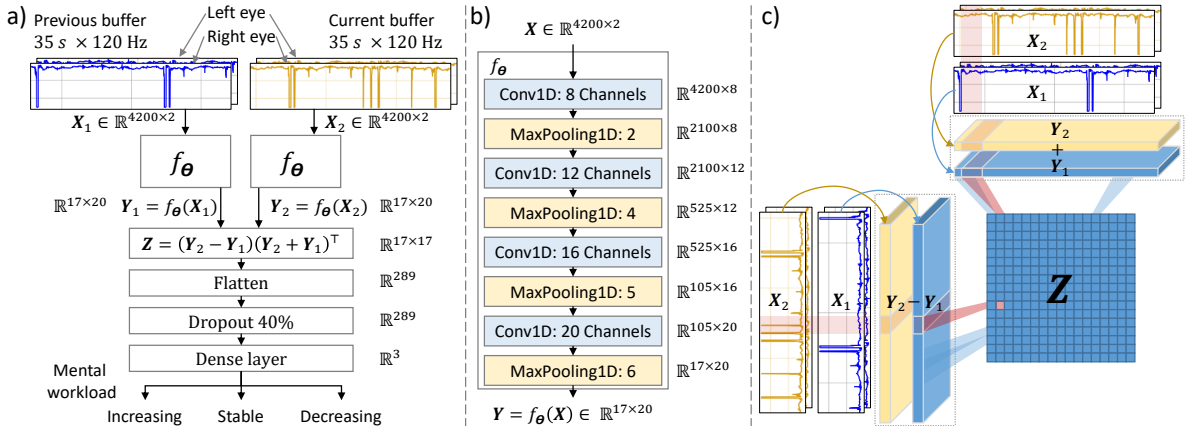

Fig. 2. The proposed solution IG-CNN: a) the overall architecture; b) $f(\boldsymbol{\theta})$ architecture; c) the features matrix $\boldsymbol{Z}$ with systematic inspection of the interaction between local segments of the previous and the current time-series

follow the original transformer encoder architecture [6]. Each block contains 4 self-attention heads and a fully connected layer. The head size is 8 while the size of the fully connected layer is 15. Residual connections and layer normalization are used as in the standard transformer encoder [6]. Each block applies a 40% dropout to fight over-fitting. The implementation of the transformer network is based on the code in [18]. The BiLSTM [8] consists in 4 bidirectional LSTM [5] layers with the dimensionality of the output space of each of them is 8. The entire sequence (not only the last output) is returned from each of these layers. Thus, the proposed design of both the transformer and the BiLSTM doesn't change the dimensionality of the input space. They are used to contextualize the representation of local segments of time series.
Finally, the the kernel size in the CNNs g_ϕ and f_θ equals 10 while the remaining configurations of these CNNs are illustrated in Fig. 1 and Fig. 2, respectively.

B. Results

Table. I shows the accuracy of the proposed solution IG-CNN compared to CNN, BiLSTM and transformer on five different modalities. IG-CNN shows the best results on all modalities except the EDA data. Pupil diameter data represents the most exploitable modality with a high accuracy of 90.5% using IG-CNN and the transformer. Fig. 3 shows the confusion matrix resulting from the evaluation of the proposed solution IG-CNN. The ambiguity between the classes "increasing" and "decreasing" is very low with only two false detection cases. Table. II shows the training time/epoch of the proposed solution compared to the considered reference networks. Just like the basic CNN, the required training time/epoch is less than 30% of the time required by the transformer or the BiLSTM.

TABLE I
THE ACCURACY (%) OF IG-CNN (PROPOSED) COMPARED TO CNN, BILSTM AND TRANSFORMER NETWORKS ON 5 MODALITIES.

Neural Network	Data source				
	EYE_P	*EYE_D*	*PPG*	*EEG*	*EDA*
CNN	86,3	74,1	67,3	72,6	45,2
BiLSTM	88,2	77,6	64,6	73,4	48,3
Transformer	**90,5**	80,2	66,9	73,8	**50,2**
IG-CNN	**90,5**	**82,5**	**77,2**	**84,4**	46,4

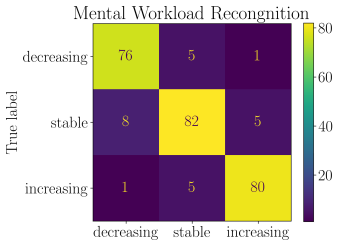

Fig. 3. The confusion matrix resulting from the evaluation of the proposed solution, IG-CNN, on relative MWL recognition using pupil diameter data.

TABLE II
THE TRAINING TIME/EPOCH (MS) OF IG-CNN (PROPOSED) COMPARED TO CNN, BILSTM AND TRANSFORMER NETWORKS ON 5 MODALITIES.

Neural Network	Data source				
	EYE_P	*EYE_D*	*PPG*	*EEG*	*EDA*
CNN	**759**	**759**	**759**	**759**	**759**
BiLSTM	3135	3135	3927	3927	3927
Transformer	2673	2706	2805	2772	2772
IG-CNN	**759**	**759**	**759**	**759**	**759**

V. CONCLUSIONS AND FUTURE WORK

An efficient solution for the classification of MWL relative changes has been proposed. The experimental results have shown the superiority of the proposed solution over classic CNN, BiLSTM and transformer networks in terms of classification accuracy. The training time is considerably smaller than the ones of BiLSTM and transformer.
In future work, the influence of inter-individual variability will be analyzed using the leave-one-subject-out cross validation strategy. In addition, the possibility of creating multi-modal features will be investigated instead of using each modality separately.

REFERENCES

[1] C. Wickens, "Multiple resources and performance prediction", Theoretical Issues in Ergonomics Science, vol. 3, pp. 159 - 177, 2002.

[2] Y. Zhou, S. Huang, Z. Xu, P. Wang, X. Wu and D. Zhang, "Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review", IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 3, pp. 799-818, Sept. 2022, doi: 10.1109/TCDS.2021.3090217.

[3] Win-Ken Beh, Yi-Hsuan Wu, An-Yeu (Andy) Wu, (2021), "MAUS: A Dataset for Mental Workload Assessmenton N-back Task Using Wearable Sensor", arXiv:2111.02561 [eess.SP].

[4] Schmiedek Florian, Lövdén Martin and Lindenberger Ulman, "A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context", Frontiers in Psychology, Volume 5 (2014), DOI=10.3389/fpsyg.2014.01475, ISSN=1664-1078.

[5] Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". Neural Computation. vol. 9 pp. 1735 – 1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276. S2CID 1915014.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", Advances in Neural Information Processing Systems, 2017, pp. 6000–6010.

[7] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770-778, 2016, doi: 10.1109/CVPR.2016.90.

[8] Graves, A., Fernández, S., Schmidhuber, J. (2005), "Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition.", "Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (eds) Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005", "Lecture Notes in Computer Science", vol 3697. Springer, Berlin, Heidelberg, 2005 https://doi.org/10.1007/11550907_126.

[9] Minghao Liu, Shengqi Ren, Siyuan Ma, Jiahui Jiao, Yizhou Chen, Zhiguang Wang and Wei Song . "Gated Transformer Networks for Multivariate Time Series Classification", 2021, arXiv:2103.14438 [cs.LG].

[10] C. Szegedy et al., "Going deeper with convolutions" "IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA", 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[11] Zerveas, George et al. "A Transformer-Based Framework for Multivariate Time Series Representation Learning", "Association for Computing Machinery, Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining", 2021, pages 2114–2124.

[12] Wang Z, Yan W, Oates T, "Time series classification from scratch with deep neural networks: a strong baseline.", "International Joint Conference on Neural Networks",2017, pp 1578-1585.

[13] Z. Zhang, Z. Pi, B. Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise", "IEEE Transactions on Biomedical Engineering", vol. 62, no. 2, pp. 522-531, February 2015, DOI: 10.1109/TBME.2014.2359372.

[14] Chollet, Francois.,"CHAPTER 6 "Deep learning for text and sequences," Deep Learning with Python.",2017, New York, NY: Manning Publications.

[15] Liyuan Liu , Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han, "On the Variance of the Adaptive Learning Rate and Beyond.",2020, "the Eighth International Conference on Learning Representations."

[16] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi., "Unsupervised Scalable Representation Learning for Multivariate Time Series.", "Advances in Neural Information Processing Systems 32, H.Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates",2019, Inc., 4650–4661.

[17] Ismail Fawaz, H., Forestier, G., Weber, J. et al., "Deep learning for time series classification: a review.", "Data Min Knowl Disc 33", 917–963, 2019. https://doi.org/10.1007/s10618-019-00619-1.

[18] Theodoros N., https://keras.io/examples/timeseries/timeseries_transformer_classification/, Timeseries classification with a Transformer model, 2021.