A Deep Learning End-to-End Approach to Mental Workload Estimation from EEG signals in flight simulation training

[Théophile Demazure[1], Alexander J Karran[1], Pierre-Majorique Léger[1], Marc Fredette[1]]
[HEC Montréal, 3000 Chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7]

Introduction

End-to-end machine learning process is gaining interest in EEG research and offers novel decoding opportunities (1). The research presented in this manuscript seeks to explore the extent to which it is possible to estimate mental workload (MW) during simulated flight training based on neurophysiological signal data using an end-to-end deep learning process. We present here our preliminary results from this exploration.

The overarching research objective is to create a deep learning model based on end-to-end methods for the offline and online assessment and estimation of MW during flight training within a high-fidelity to cockpit scale "fast jet" simulator. Our aims are twofold:

1. Explore a mental workload classifier that achieves high classification performance, displays fast and stable convergence, and increases its computational efficiency for potential online usage.
2. Estimate mental workload during a flight simulator training task and replicate past empirical findings related to the link between task complexity, mental workload, and performance.

A review of MW estimation using deep learning approaches and EEG signal data revealed that fully utilizing end-to-end processes with no manual feature engineering is still rare. We benchmarked several deep learning models for multivariate time-series classification and selected the two best-performing architectures. Informed by the ML and neuroscience literature, we iterated model design choices specific to EEG signal data and systematically assessed those choices.

Methods

Participants

Eleven participants (mean age = 34.56, SD = 8.97), all novice pilots, participated in the study. Ethical approval was conducted by internal review by our aerospace partner, and participants were recruited from within the company. The study gained additional ethical approval from the HEC Montréal REB.

Experimental Task and Environment

Participants performed two tasks: a synthetic task consisting of an *n*-back, designed to manipulate mental workload, and an ecologically valid flight task designed to induce different levels of mental workload that mirror *n*-back difficulty levels through manipulating manoeuvre difficulty. In this case, four low, four moderate and three high complexity manoeuvres.

For the *n*-back task, we used a letter stimulus and identity recognition (i.e., the same letter as presented *n* trials previously) design. To obtain a granular classification of MW, we incremented the number of iterations of *n* from 0 to 3 with 40 trials for each. Participants were allowed ten practise trials before beginning the task. To assess participant subjective assessment of mental workload throughout the flight task and as a manipulation check, participants were asked to complete the NASA RAW-TLX.

Data processing

EEG signal was acquired using a 32 channel BrainVision headset following the standard 10–20 montage. Signal was filtered using a bandpass 1-40 Hz Butterworth 2nd order IIR filter. Artefacts were removed manually using blind source separation by independent component analysis (extended infomax). The filtered EEG signal data were then downsampled to 500 Hz (from 1000hz), then segmented into *n*=160 epochs of 3s (-100ms to 2900ms) per participant for a total of 1440 epochs to create the training set, very low and low, medium and high *n*-back levels were merged to create a binary classification problem. For the flight task, available data was chunked at 3s for classification. A window size of 3s provides enough data to investigate the effect of window size upon model performance. Data for two participants were dropped after failing quality assessment.

Model Benchmarking & Validation

We selected six models for testing and comparison: a Multi-Layer Perceptron (MLP), a Fully Convolutional Neural Network (FCN), Time Convolutional Neural Network (T-CNN), a Multi-Channel Deep Convolutional Neural Network (MCDCNN), a Multi-scale Convolutional Neural network (MCNN), and a Residual Network (ResNet). Models were trained five times using 5-fold cross-validation to create an accuracy baseline for subject-dependent classification; then, further training modifications were iterated through model design choices such as drop out for hidden layers implemented to reduce overfitting and then applied to the flight data. The models produced an average of 187, 288 and 388 classifications for low, moderate, and high manoeuvre complexity blocks.

Validation was performed in three ways 1.) for the flight task through a single factor, within-subject repeated-measures ANOVA (Bonferroni corrected) to determine the variance between perceived mental workload per manoeuvre block, pairwise t-tests showed a significant effect of manoeuvre complexity upon perceived MW, $F(2,20) = 7.995$, $p = 0.003$), 2.) An analysis of variance to determine the relationship between performance and manoeuvre complexity showed a significative inverse relationship where performance decreased as flight manoeuvre complexity increased $F(10,80) = 3.69$, $p = .001$, CI = [.07,.37] 3.) Estimations of low, moderate, and high MW were determined through majority vote from binary classification output (2) logistic regression with a random intercept model and a fixed effect for task complexity was performed upon these data, which reported a significant negative relationship between performance and estimated MW for both selected models FCN ($p<0.01$, r^2 0.117) and ResNet ($p<0.01$. r^2 0.099) as manoeuvre complexity increased in line with expected responses.

Table 1 Tested design choices and justification

Design Choice	Tested Design Choices	Justification
Optimizers	Adam, SGD, Adadelta, Nadam	Rapid convergence and stability Features identification
Training window	Full trial windows, half trial windows	The varying length of windows can impact the performance in EEG learning tasks
Activation function	ReLU and eLU	Relu as shown to be faster in the hidden layers for deep neural networks, eLU shown to perform well with residual gates in EEG learning tasks
Dropout	From 0.5 to 0.2 on hidden layers	Avoid coadaptation Force the model to learn generalizable features instead of focusing on highly predictive one Performance

Results

The two selected classifier models achieved an average F1 performance of μ=92.6 (σ .056) and μ=91.2 (σ =.109) for FCN and ResNet when trained, respectively. A statistical comparison of all models using a Wilcoxon signed-rank test (Benjamini-Hochberg corrected) indicated that both FCN and ResNet presented significantly superior classification accuracy; no significant difference in accuracy between the baseline and iterated (.2 dropout) models were reported between FCN and ResNet. Thus, we retained both models to systematically benchmark the impact of further design choices in line with our research aims (Table 1.).

Feature-level assessment

To assess the acceptability of features learned by the model, we used Integrated Gradients (IG). IG is a backpropagation-based method that evaluates how a model's input features affect its predictions (3) and has been shown to be an appropriate method for capturing global nonlinear effects and cross interactions between different features (4). To make IG understandable, we computed the average marginal contribution of all EEG channels towards an estimation of high workload per trial. We then visualized these gradients to illustrate positive contributions, where red shows the positive influence of a channel's features towards the prediction of a high workload level, and blue the negative contribution that influences the prediction away from the expected value.

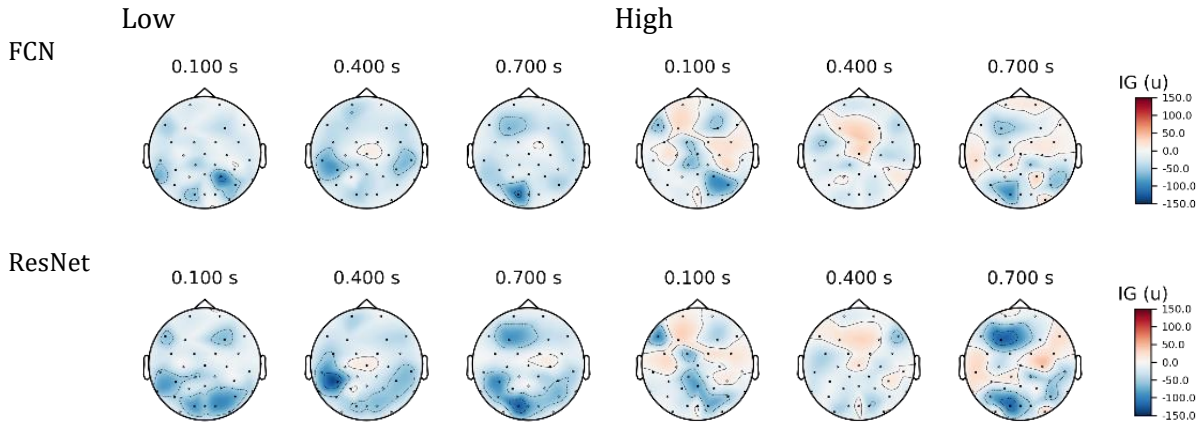

Figure 1. Integrated Gradient grand average for each condition, computed using the IG values averaged for all participants between the multiple folds. Note that these visualizations oversimplify what features a model has learned and do not represent the complexity of the models' behaviour.

Discussion

These preliminary findings are very promising, and we foresee that future assessments will only strengthen the results. We evaluated the learned features of the models through feature attribution and contribution to assess their validity for a neurophysiological inference of MW. We posit that end-to-end deep learning analysis can offer a powerful complementary alternative to traditional approaches for mental state estimation through automated feature discovery. However, this method creates a new set of challenges to be addressed. When applying deep learning to learn features and estimate mental state directly from signal data, a model will utilize any discriminant feature regardless of whether that feature is related to the target cognitive state. Our future work will assess how these predictive artefacts may confound and affect a model's performance and validity (either positively or negatively), depending on the task.

References:

1. Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J. Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering. 2019;16(5).
2. Karran AJ, Fairclough SH, Gilleade K. A framework for psychophysiological classification within a cultural heritage context using interest. ACM Transactions on Computer-Human Interaction (TOCHI). 2015;21(6):1-19.
3. Sundararajan M, Taly A, Yan Q, editors. Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning-Volume 70; 2017: JMLR. org.
4. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:171106104. 2017.