2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Learning Generalized Representations of EEG between Multiple Cognitive Attention Tasks

Yi Ding*, Nigel Wei Jun Ang*, Aung Aung Phyo Wai, Cuntai Guan**

Abstract— **Attention can be measured by different types of cognitive tasks, such as Stroop, Eriksen Flanker, and Psychomotor Vigilance Task (PVT). Despite the differing content of the three cognitive tasks, they all require the use of visual attention. To learn the generalized representations from the electroencephalogram (EEG) of different cognitive attention tasks, extensive intra and inter-task attention classification experiments were conducted on three types of attention task data using SVM, EEGNet, and TSception. With cross-validation in intra-task experiments, TSception has significantly higher classification accuracies than other methods, achieving 82.48%, 88.22%, and 87.31% for Stroop, Flanker, and PVT tests respectively. For inter-task experiments, deep learning methods showed superior performance over SVM with most of the accuracy drops not being significant. Our experiments indicate that there is common knowledge that exists across cognitive attention tasks, and deep learning methods can learn generalized representations better.**

I. Introduction

Brain-Computer Interface (BCI) systems were initially used for communication and control in paralyzed patients, but have seen use in the form of cognitive training and tests to model and evaluate cognitive states of healthy people [1]. EEG BCI experiments generally consist of a few electrodes placed on a subject's head. This allows for lower cost, non-invasive EEG experiments and has thus become an increasingly preferred choice in passive BCI research domains and attention-focused research [2] [3].

Attention is paramount in many daily tasks that require attention and focus, such as studying, operating machinery, etc. [4]. Attention training using EEG BCI systems has been shown to improve ADHD symptoms [3]. It can be broadly divided into four types: selective attention, sustained attention, divided attention, and executive attention [5]. Selective attention plays an important role in human intelligence as it requires the ability to focus on desired information while avoiding distractions [4]. Sustained attention is highly required during tasks like driving vehicles, monitoring surveillance footage, or concentrating during study sessions [6] [7]. Stroop [8], Flanker [9], and psychomotor vigilance task (PVT) tests [10] are typically used to measure the attention state of human beings. Stroop tasks measure selective attention by instructing the subject to name the color of words [8]. Flanker tasks measure selective attention by instructing test subjects to respond to a target stimulus surrounded by non-target stimuli [11]. PVT measures the response speed of certain visual stimuli [10]. Despite the differing content of the three cognitive tasks, they all require the use of visual attention.

Deep Learning methods have achieved superior accuracy over Machine Learning models when applied to classification tasks on EEG data [12] [13] [14]. Robinson et al. [15] developed an EEG BCI system for hand-motor imagery decoding with CNNs. Lawhern et al. [16] proposed EEGNet, capable of capturing spatial and temporal information through convolutional kernels. Ding et al. [12] developed TSception; a novel deep learning framework, achieving high accuracy of 86.03% and outperforming previous methods for emotion state classification.

In order to learn the generalized representations among different mental tasks, we conduct both intra-task and inter-task attention state classification experiments on the attention dataset proposed in [5] which contains EEG data of Stroop, Flanker, and psychomotor vigilance task (PVT) tasks. SVM, EEGNet, and TSception are used in our experiments. Section II describes the details of the dataset, experiment settings and evaluation process. The results and analysis are provided in section III. Finally the we discuss and conclude the paper in section VI.

II. Materials and methods

A. Dataset and Pre-processing

The dataset used in this work was proposed in [5]. 10 healthy adults participated in the experiments. The EEG data were recorded using a wearable EEG MUSE headband and eye gazes were captured by a desktop eye tracker. Four EEG electrodes (Tp9, AF7, AF8, Tp10) are used in the dataset. Each subject is required to perform three types of mental tasks, Stroop, Flanker, and PVT. There are two classes in each type of mental tasks, attention and inattention trials. For each attention trial, subjects are required to repeat the desired mental task 10-13 times. For inattention trial, the design is the same for 3 types of mental attention tasks, which introduces the subject not to focus by slowly looking around on the black screen. Each task contains three sessions. In each session, there are three blocks which contains one attention trial and one inattention trial. Each trial lasts for 30s. Hence, there are 18 trials (9 attention and 9 inattention) in total for each mental attention task.

First, the data is band-passed using a zero-phase fifth-order IIR band-pass filters of 0.3-45Hz to remove low and high-frequency noise. Average filters are then applied to remove artifacts as [5]. Each trial's data is segmented using sliding

Yi Ding, Nigel Wei Jun Ang, Aung Aung Phyo Wai and Cuntai Guan are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. 50 Nanyang Ave, Singapore 639798. * These authors contributed equally. ** Cuntai Guan is the Corresponding Author. Email: ctguan@ntu.edu.sg

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

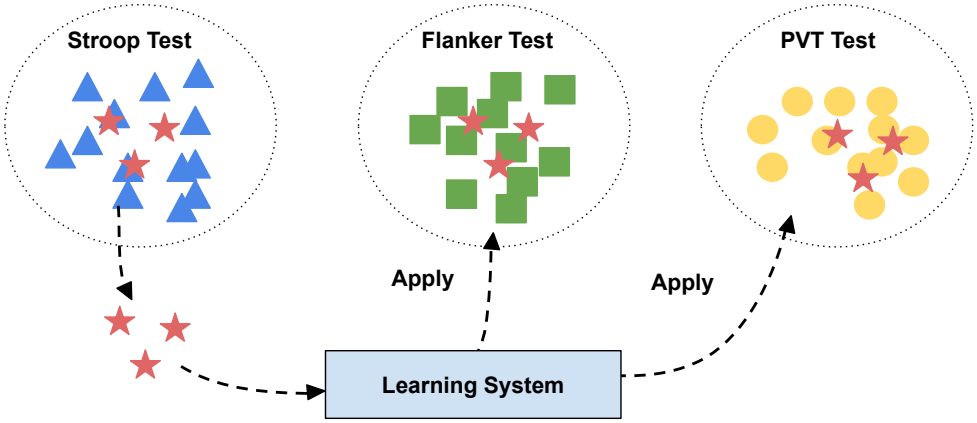

Fig. 1. Learning the common hidden knowledge from one task and apply the learned knowledge to other related tasks. Three mental attention tasks are different in format but they all require visual attention during the task. We hold the hypothesis that the learning system can learn such common hidden knowledge from each task which can be applied to other attention tasks. Learning from Stroop data is shown just for illustration. The study in this paper is done for all inter-task scenarios.

windows of lengths 1s, 2s, and 4s with a moving step of 100ms. For SVM, features are further extracted. Differential entropy features of four frequency bands are extracted as the input of SVM as [12]. The SVM is used as the baseline, since it is the most commonly used machine learning methods in BCI [5]. For the deep learning methods, they can learn the classification-related representations from EEG data via the convolutional layers. Hence the EEG segments are fed into the EEGNet and TSception directly.

B. Intra-task Attention Classification

For each subject (subject-specific experiment), intra-task experiments are conducted to evaluate the classification performance of SVM, EEGNet, and TSception on the binary classification (attention vs. inattention) of each type of mental tasks. For each task's data, leave one block out cross-validation [5] is utilized to evaluate three methods. Each time one block's data are selected as testing data, the remaining data are used as training data. Among the training data, 20% are selected as validation data and the remaining 80% are used as training data. This process is repeated until every block's data are used as testing data once. The average classification accuracy of leave one block out cross validation is used as the classification performance on each subject. The final mean classification accuracy of all subjects is reported for each task. Let $\mathbf{X}_i^{\mathcal{T}} \in \mathbb{R}^{c\times l}, \mathbf{Y}_i^{\mathcal{T}} \in \mathbb{R}, i \in [1, ..., n]$, be the data and label of task $\mathcal{T}$, where c is the number of EEG channels, l is the length of each EEG segment, and n is the total number of EEG segments. For deep learning methods, the optimization problem of the intra-task experiment is to find the classifier $\Phi(\cdot)$ parameterized by Θ which can minimize the below loss function as [12]:

$$\mathcal{L}^{\mathcal{T}} = \sum_{i=1}^{n} \mathcal{L}_{CE}(\mathbf{Y}_i^{\mathcal{T}}, \Phi(\mathbf{X}_i^{\mathcal{T}}, \Theta)) + \lambda \mathcal{L}_1(\Theta) \qquad (1)$$

where the $\mathcal{L}_{CE}$ is the cross-entropy loss, $\mathcal{L}_1$ is the L1 regularization term, λ is the L1 regularization coefficient.

C. Inter-task Attention Classification

Stroop, Flanker, and PVT tests are different mental attention tasks. Although they are different in experiment designs, all of them require the subjects to have high visual attention during the tasks. We hold the hypothesis that deep learning methods can better learn this common information as the generalized representations. Hence, the inter-task experiments are conducted to evaluate the inter-task classification performance of these three methods.

For each subject (subject-specific experiment), inter-task experiments are conducted. During the experiment, one particular task is selected as the training task, the other two are used as testing tasks. For the training task, two blocks are used as validation data, the other seven blocks are used as training data. During the training process, **Eq.** 1 is utilized to guide the network training. Early stopping is adopted to save the model which achieves the best accuracy on validation data. The saved model will be evaluated on the testing tasks separately. Given $\mathbf{X}_i^{\mathcal{T}_j} \in \mathbb{R}^{c\times l}, \mathbf{Y}_i^{\mathcal{T}_j} \in \mathbb{R}, i \in [1, ..., n], j \in [1, 2, 3]$ as the data and label of task $\mathcal{T}_j$. $\mathcal{T} \in$ [stp, flk, pvt], where stp = Stroop test, flk = Flanker test, pvt = PVT. Let $\Phi^{\mathcal{T}_j}(\cdot)$ denote the classifier trained on task $\mathcal{T}_j$. For each task, we separately apply the $\Phi^{\mathcal{T}_j}(\cdot)$ on task $\mathcal{T}_m$, where $m \in [1, 2, 3], m \neq j$, to calculate the cross-task accuracy. This process is repeated until all the three tasks are used as the training task once. The average classification accuracy of all the subjects is reported for each inter-task experiment.

D. Implementation Details

For SVM, RBF kernel with C being 2 and γ being 0.5 is used. For the deep learning methods, PyTorch library is used. The experiment is run on an Ubuntu 18.04 machine with

TABLE I

MEAN ACCURACY FOR INTRA-TASK ATTENTION CLASSIFICATION USING LEAVE ONE BLOCK OUT CROSS-VALIDATION

Tasks	Methods	Segment length		
		1s	2s	4s
Stroop	SVM	66.56%***	69.31%***	71.61%**
	EEGNet	68.48%**	75.03%*	78.21%*
	TSception	**72.23%**	**77.99%**	**82.48%**
Flanker	SVM	74.53%*	76.38%*	76.39%**
	EEGNet	71.57%**	77.10%	77.26%*
	TSception	**78.46%**	**83.05%**	**88.22%**
PVT	SVM	73.86%*	76.32%**	78.09%**
	EEGNet	74.07%	80.67%	83.38%
	TSception	**77.43%**	**82.13%**	**87.31%**

$*$:p-value < 0.05; $**$:p-value < 0.01; $***$:p-value < 0.001.
The p-value is between TSception and other methods since TSception achieves the highest accuracy (bold font in the table data) in all settings of all the tasks.

a Tesla V100 GPU. The hyper-parameters of EEGNet and TSception are set to be the suggested ones in their original paper. The batch size is set as 64, the Early Stopping patience is set as 5, which means the training process will cease once the accuracy on the validation set does not increase for 5 epochs. L1 regularization is used as [12]. The L1 regularization coefficient is 1e-6. The maximum training epoch is set as 100, while the dropout rate is set as 0.5. Adam optimizer is adopted with the initial learning rate being 1e-3.

III. RESULTS AND ANALYSIS

A. Intra-task Experiment Results

The classification results of the three methods in intra-task experiments are shown in **TABLE I**. Paired T-test is utilized to do the statistical analysis of the results. TSception achieves significantly higher accuracies than other methods in all the experiments. EEGNet achieves second place in most of the experiments except Flanker test in which SVM is slightly higher than EEGNet when the segment lengths is 1s. Generally, deep learning methods show superior classification ability to SVM. It is also obvious that the mean accuracy of all three methods increases as the segment length increases. Among the three tasks, the best classification accuracies (achieved by TSception) in three types of segment lengths are all higher than the other two models, being 78.46% for 1s, 83.05% for 2s, and 88.22% for 4s. However the best classification results of all three segment lengths in Stroop test experiments are relatively lower than others, being 72.23% for 1s, 77.99% for 2s, and 82.48% for 4s respectively. Besides, the overall performances of all three methods for Stroop test experiments were lower than Flanker test and PVT ones, indicating Flanker test and PVT have better attention inducing ability.

B. Inter-task Experiments Results

Although the three cognitive tasks are different, visual attention is required during all the tasks. We hold the hypothesis that deep learning methods can learn the common hidden knowledge to generalize better across tasks. The inter-task experiment results are shown separately for three different machine/deep learning methods first in **TABLE II - IV** to compare the accuracy changes between training on the task's own data and training on other tasks' data. Then the accuracy comparisons among different methods for each inter-task experiment are shown in **Fig. 2**. Paired T-test is used for statistical analysis.

According to **TABLE II**, all the accuracies of inter-experiments drop compared to the intra-task ones using SVM. The accuracy drops of most inter-task experiments are not statistically significant ($p > 0.05$), except when training the SVM on Stroop data then testing on Flanker ($p = 0.006$ for 1s; $p = 0.019$ for 2s; $p = 0.085$ for 4s) and PVT ($p = 0.006$ for 1s; $p = 0.001$ for 2s; $p < 0.001$ for 4s). Training SVM on Flanker test data gives the minimum overall drop with the maximum accuracies being 65.49% (1s), 66.51% (2s) and 68.62% (4s) for testing on Stroop and 71.48% (1s), 73.30% (2s) and 72.63% (4s) for testing on PVT. The overall trend remains that the longer the segment length, the higher the accuracy.

TABLE III shows the cross tasks performance of EEGNet. In most of the experiments, the accuracy drops are not significant ($p > 0.05$). The minimum drops are observed when EEGNet is trained on PVT data and tested on Stroop data. The accuracy decreased by 1.37% when segment length is 1s ($p = 0.642$); for 2s segments it drops from 75.03% to 72.06% ($p = 0.317$) and the smallest drop is 0.44% ($p = 0.914$) for 4s segment setting. The accuracies even increase when training on PVT task data and tested on Flanker task data, especially when the segment length is 4s, which has 9.5% improvement compared with the intra-task results. The best overall performances of EEGNet are seen in the experiments using PVT data as training data. Like the SVM, the accuracy will increase as the data segment length increases.

The accuracies of TSception for inter-task experiment are shown in **TABLE IV**. Although the drops of accuracies using TSception are relatively larger than SVM and EEGNet, most of the accuracy drops are not significant. When the model trained on PVT data is used to classify the Flanker test data, smallest drops are observed in all three types of segments with the drops being 1.68% for 1s ($p = 0.390$), 1.94% for 2s ($p = 0.253$) and 3.12% for 4s ($p = 0.144$). The

TABLE II

MEAN ACCURACY FOR INTER-TASK ATTENTION CLASSIFICATION USING SVM

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	66.56%	67.06%	66.66%
Flanker	65.49%	74.53%	71.48%
PVT	64.82%	71.41%	73.86%

(a) Segment length=1s

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	69.31%	69.26%	68.54%
Flanker	66.51%	76.38%	73.30%
PVT	66.17%	72.66%	76.32%

(b) Segment length=2s

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	71.61%	70.65%	69.96%
Flanker	68.62%	76.39%	72.63%
PVT	67.82%	72.13%	78.09%

(c) Segment length=4s

TABLE III

MEAN ACCURACY FOR INTER-TASK ATTENTION CLASSIFICATION USING EEGNET

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	68.48%	70.71%	67.64%
Flanker	66.17%	71.57%	70.38%
PVT	67.11%	71.80%	74.07%

(a) Segment length=1s

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	75.03%	75.83%	73.31%
Flanker	71.61%	77.10%	74.98%
PVT	72.06%	78.18%	80.67%

(b) Segment length=2s

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	78.21%	78.25%	74.68%
Flanker	74.48%	77.26%	76.60%
PVT	77.77%	86.76%	83.38%

(c) Segment length=4s

TABLE IV

MEAN ACCURACY FOR INTER-TASK ATTENTION CLASSIFICATION USING TSCEPTION

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	72.23%	73.64%	72.41%
Flanker	67.62%	78.46%	72.03%
PVT	68.87%	76.78%	77.43%

(a) Segment length=1s

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	77.99%	78.84%	76.19%
Flanker	73.33%	83.05%	76.28%
PVT	74.38%	81.11%	82.13%

(b) Segment length=2s

	Test on		
Train on	Stroop	Flanker	PVT
Stroop	82.48%	83.15%	80.38%
Flanker	79.52%	88.22%	83.36%
PVT	79.47%	85.10%	87.31%

(c) Segment length=4s

same thing happens if the model is trained on Stroop data and tested on PVT data with relative larger drops than the ones using Flanker data as testing data. Training TSception on PVT data gives the best overall performance with the maximum accuracies being 68.87% (1s), 74.38% (2s) and 79.47% (4s) for testing on Stroop and 76.78% (1s), 81.11% (2s) and 85.10% (4s) for testing on Flanker. The accuracies are positively correlated to the segment lengths as well for TSception.

The comparisons between different machine/deep learning methods are shown in **Fig. 2**. TSception still achieves the best classification in inter-task experiments, except when training on PVT and testing on Flanker experiment with 4s segment length. From the bar chart, both deep learning methods are better than SVM, and the improvements increase with the increment of segment lengths.

IV. DISCUSSION AND CONCLUSION

Attention can be measured by cognitive attention tasks. Stroop task asks the subject to name the color of words to measure the selective mental attention. Flanker task requires subject to response to the certain stimulus only when the target stimulus is surrounded by non-target ones. PVT measures the response speed to certain visual stimulus. Although the three mental tasks are different in content, they all required visual attention during the tasks. This visual attention can be regarded as the common hidden knowledge among the three tasks. We hold the hypothesis that deep learning methods can learn this hidden knowledge and extract a much more generalized representation for classification tasks across different mental attention tests. To evaluate our hypothesis, both intra and inter-task attention classification experiments were conducted using three modern machine/deep learning methods.

In the intra-task experiment, the classification performance of all three methods were evaluated. According to the results, both deep learning methods achieved higher accuracies than SVM, showing the better feature learning ability. Among the two deep learning methods, TSception achieves significantly higher accuracy than EEGNet. In the inter-task experiments, a model was trained on one mental task's data and evaluated on other tasks. From **Table** IV (c), which lists down the accuracy using 4s data and achieved the highest performance among all 3 segment lengths, it can be observed that the intra-task cases achieve the highest accuracy, but the inter-task cases only suffer a small drop in accuracy. Therefore, we hypothesize that there might be common features in EEG which may represent the manifestation of the common attention/inattention process in the brain. Besides, both deep learning methods achieve higher classification accuracies than SVM in inter-tasks experiments, indicating deep learning methods have better learning ability for the common hidden knowledge across different cognitive attention tasks. For both EEGNet and TSception, the best overall results were

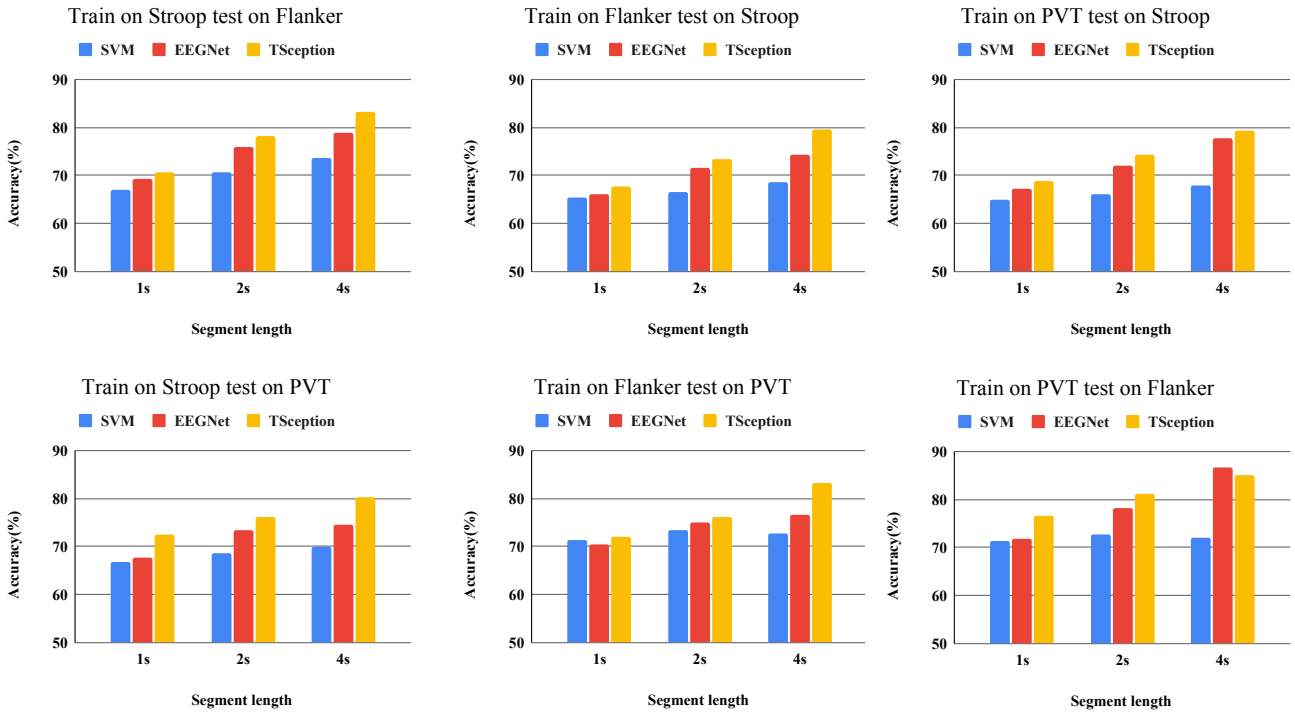

Fig. 2. Performance comparison between SVM, EEGNet and TSception in inter-task experiments.

observed when they were trained on PVT task data while the worst ones are in the experiments where Stroop data were used as training data, indicating less common knowledge between Stroop and other two tasks. In the future, efforts will be given to design new training strategies to better learn the common hidden knowledge among different cognitive attention tests.

ACKNOWLEDGMENT

This work was partially supported by the RIE2020 AME Programmatic Fund, Singapore (No. A20G8b0102).

REFERENCES

[1] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *Journal of Neural Engineering*, vol. 16, no. 1, p. 011001, Jan 2019.

[2] P. Aricò, G. Borghini, G. Di Flumeri, N. Sciaraffa, A. Colosimo, and F. Babiloni, "Passive BCI in operational environments: Insights, recent advances, and future trends," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1431–1436, 2017.

[3] C. G. Lim, T. S. Lee, C. Guan, D. S. S. Fung, Y. Zhao, S. S. W. Teng, H. Zhang, and K. R. R. Krishnan, "A Brain-Computer Interface Based Attention Training Program for Treating Attention Deficit Hyperactivity Disorder," *PLoS ONE*, vol. 7, no. 10, 2012.

[4] C. M. JUNG, J. M. RONDA, C. A. CZEISLER, and K. P. WRIGHT JR, "Comparison of sustained attention assessed by auditory and visual psychomotor vigilance tasks prior to and during sleep deprivation," *Journal of Sleep Research*, vol. 20, no. 2, pp. 348–355, 2011.

[5] A. A. Phyo Wai, M. Dou, and C. Guan, "Generalizability of EEG-based mental attention modeling with multiple cognitive tasks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 2959–2962.

[6] R. F. Helfrich, I. C. Fiebelkorn, S. M. Szczepanski, J. J. Lin, J. Parvizi, R. T. Knight, and S. Kastner, "Neural mechanisms of sustained attention are rhythmic," *Neuron*, vol. 99, no. 4, pp. 854–865.e5, 2018.

[7] Z. Guo, Y. Pan, G. Zhao, S. Cao, and J. Zhang, "Detection of driver vigilance level using eeg signals and driving contexts," *IEEE Transactions on Reliability*, vol. 67, no. 1, pp. 370–380, 2018.

[8] F. Scarpina and S. Tagini, "The stroop color and word test," *Frontiers in Psychology*, vol. 8, p. 557, 2017.

[9] T. J. McDermott, A. I. Wiesman, A. L. Proskovec, E. Heinrichs-Graham, and T. W. Wilson, "Spatiotemporal oscillatory dynamics of visual selective attention during a flanker task," *NeuroImage*, vol. 156, pp. 277–285, 2017.

[10] C. Formentin, M. De Rui, M. Zoncapè, S. Ceccato, L. Zarantonello, M. Senzolo, P. Burra, P. Angeli, P. Amodio, and S. Montagnese, "The psychomotor vigilance task: Role in the diagnosis of hepatic encephalopathy and relationship with driving ability," *Journal of Hepatology*, vol. 70, no. 4, pp. 648–657, 2019.

[11] M. J. Imburgio, I. Banica, K. E. Hill, A. Weinberg, D. Foti, and A. MacNamara, "Establishing norms for error-related brain activity during the arrow flanker task among young adults," *NeuroImage*, vol. 213, p. 116694, 2020.

[12] Y. Ding, N. Robinson, Q. Zeng, D. Chen, A. A. Phyo Wai, T. S. Lee, and C. Guan, "TSception:a deep learning framework for emotion detection using EEG," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[13] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, apr 2019.

[14] 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.

[15] N. Robinson, S. Lee, and C. Guan, "EEG representation in deep convolutional neural networks for classification of motor imagery," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 2019, pp. 1322–1326.

[16] 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.