Generalization of Deep Neural Networks for EEG Data Analysis

by

Yurui Ming

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Doctor of Philosophy

Supervisor: Prof. Chin-Teng Lin
Co-supervisor: Dr. Mukesh Prasad

Computational Intelligence and Brain Computer Interfaces Lab (CIBCI)
Centre for Artificial Intelligence (CAI)
School of Computer Science
Faculty of Engineering and Information Technology (FEIT)
University of Technology Sydney (UTS)

Sydney, Australia
2020

Certificate of Authorship and Originality

This is to certify that this thesis is submitted in fulfilment of the requirements for award of Doctoral of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney.

This thesis is wholly the author's own work unless otherwise referenced or acknowledged. In addition, the author certifies that all information sources and literature used are indicated in this thesis. This document has not been previously submitted for qualification of a degree in any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

15/09/2020

Abstract

Electroencephalography (EEG) facilitates the neuroscientific research and applications by virtue of its properties such as non-invasion, affordability, mobility, etc. However, challenges including high artefacts pending, intra- and cross-subject variance, limited data availability, etc., pose the difficulty in reaching solid conclusions. This thesis explores how to utilize and generalize deep neural networks (DNN), which have set new performance records in various fields, to analyze EEG data to mitigate these challenges in reaching enlightening conclusions.

This thesis brings the comprehension of research goals by the introduction of EEG and DNN background. It reviews the conventional EEG signal processing methods and highlights the challenges of EEG data analysis. As part of this work, EEG datasets from three stereotypical brain-computer interface (BCI) experiments are described in detail to assess the proposed methods by benchmarking. To illustrate the DNN centered methodology for addressing divergent challenges of EEG data analysis, a research map is compassed to show respective contributions in fulfilling the following specific goals: (1) Selection of appropriate DNN structures targeting EEG data captured during different BCI experiments. (2) Solutions to address the intra- and cross-subject variance of EEG data. (3) Utilization of brain-inspired computation such as memory network to improve the performance of processing EEG data. (4) Exploration of new computation paradigm, i.e., reinforcement learning (RL), to relieve the noise label challenge and to improve data utilization.

By a series of published and in-preparation papers, this thesis demonstrates different achievements corresponding to the set goals: (1) Investigation of the computation traits of neural network structures and revelation of their effectiveness in EEG signal processing. The designed recurrent residual network (RRN), which is based on the recurrent structure, residual structures, etc., achieves the highest classification accuracy and provides coherent evidence and interpretation to the efficacy of conventional hand-crafted filters. (2) Invention of adversarial method in light of the domain adaptation (DA) and generative adversarial network (GAN) to address inter- and cross-subject variance. The proposed subject adaptation network (SAN), which borrows the philosophy of GAN but works in different ways, shows promising results among EEG sample clustering, sample-of-interest selection, EEG data alignment, etc. (3) Systematic study of memory networks and proposal of memory module based on self-organized maps (SOM). Work on a stacked version of differentiable neural computer (DNC), reveals that EEG features buried in the endurance experiment can be more effectively harvested by augmenting memory and consequently boost the performance. SOM-based memory network demonstrates its capability in reducing network complexity. (4) Implementation of reinforcement learning (RL) for EEG data analysis to relieve the noisy label challenge and to improve EEG data utilization. Instantiation of RL framework such as deep Q-network (DQN) demonstrates its feasibility and practicability for certain BCI experiments. Generally, through this sequence of work and papers, this thesis contributes from different aspects that well advance the EEG data analysis via DNN.

Acknowledgements

First and foremost, the author thanks Professor Chin-Teng (CT) Lin for his great support and supervision during the whole PhD candidature period. His insight, sharpness and brevity in delivering the guidance to the research is evaluable for the author's work. The author also thanks Professor CT for offering the opportunities involving in other academic activities, which are necessary to train independent researchers.

Helps and supports from lab members and collaborators such as Dr. Mukesh Prasad, Dr. Weiping Ding, Dr. Dongrui Wu, Dr. Yu-Kai Wang, et al., are highly appreciated and it is also manifest from the published/revised/submitted/preparing papers enlisted as co-authors. The author also expresses the cherished friendship with lab mates such as Fred, Carlos, Thong, Howe, Jia, et al. developed over these years.

In addition, the author thanks Dr. Avinash K. Singh, Dr. Sonny Ce, et al., for their frequent discussions about different aspects of the research, and the administration team in the School of Computer Science such as Margot, Janet and Caitlin for administrating affairs, and other PhD students in CIBCI lab for daily communications. The author also pays attributes to the following people for a variety of help hard to be categorized but contributing to an enjoyable and beneficial experience during this period: Dr. Jiaochao Yao, Dr. Xian Tao, Xiaowei, et al.

Lastly, the author appreciates the understanding and sacrifices made from family members, and well acknowledged that this work is supported by UTS President's (UTSP) Scholarship and UTS International Research Scholarship (UTS IRS) and in part by the Australian Research Council (ARC) under discovery grant DP180100670 and DP180100656.

Publications

Conference Papers

- Yurui Ming, Yu-Kai Wang, Mukesh Prasad, Dongrui Wu and Chin-Teng Lin, "Sustained Attention Driving Task Analysis based on Recurrent Residual Neural Network using EEG Data," IEEE International Conference on Fuzzy Systems, 2018.

Journal Papers

- Jung-Tai King, Mukesh Prasad, Tsen Tsai, Yurui Ming and Chin-Teng Lin, "Influence of Time Pressure on Inhibitory Brain Control During Emergency Driving," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018. https://doi.org/10.1109/TSMC.2018.2850323

- Yurui Ming, Danilo Pelusi, Chieh-Ning Fang, Mukesh Prasad, Yu-Kai Wang, Dongrui Wu and Chin-Teng Lin, "EEG data analysis with stacked differentiable neural computers," Neural Computing and Applications, 2018. https://doi.org/10.1007/s00521-018-3879-1

- Yurui Ming, Chin-Teng Lin, Stephen D. Bartlett, Wei-Wei Zhang, "Quantum topology identification with deep neural networks and quantum walks," npj Computational Materials, Aug. 2019. https://doi.org/10.1007/s11467-019-0918-z

- Yurui Ming, Danilo Pelusi, Weiping Ding, Yu-Kai Wang, Mukesh Prasad, Dongrui Wu, Chin-Teng Lin, "Subject Adaptation Network for EEG Data Analysis," Applied Soft Computing, Sep. 2019. https://doi.org/10.1016/j.asoc.2019.105689

- Yurui Ming, Dongrui Wu, Yu-Kai Wang, Yuhui Shi, Chin-Teng Lin, "EEG-based Drowsiness Estimation for Safety Driving using Deep Q-Learning". IEEE Transactions on Emerging Topics in Computational Intelligence, 2020. https://doi.org/10.1109/TETCI.2020.2997031

- Yurui Ming, Weiping Ding, Yu-Kai Wang, Chin-Teng Lin, " Memory Augmented Convolutional Neural Network and Its Application," In Preparation.

List of Abbreviations

ANN ····· Artificial Neural Networks
BCI····· Brain Computer Interface
CNN ····· Convolutional Neural Networks
DA ····· Domain Adaptation
DL····· Deep Learning
DNN ····· Deep Neural Networks
DNC ····· Differentiable Neural Computer
DQN ····· Deep Q-Network
DRL····· Deep Reinforcement Learning
EEG ····· Electroencephalography
GAN ····· Generative Adversarial Networks
LSTM····· Long Short-term Memory
ML ····· Machine Learning
MN ····· Memory Network
RL ····· Reinforcement Learning
RNN ····· Recurrent Neural Networks
SAN ····· Subject Adaptation Network
SL ····· Supervised Learning
SOM ····· Self-organizing Maps

Content

1. Introduction

1.1 EEG Brain Imaging

Brain is the source of human intelligence and emotions [1]. A better understanding of its functionalities can benefit people's professional and private life. In addition, increased knowledge gained from brain research can aid doctors in developing treatments for neurodegenerative diseases and help engineers to design more elegant brain-computer interface (BCI) based applications [2]. Furthermore, brain also inspires the artificial neural networks (ANN), which bear the current hottest machine learning (ML) research, i.e., deep learning (DL) or deep neural networks (DNN) [3].

However, brain research and application are of great challenges and involving multidisciplinary approaches from cell biology, physiology, anatomy, clinical practice, to behavioural science [4]. During last decades, great progress is made on areas that greatly benefit the society, and brain research and applications stay on the frontiers of modern multidiscipline [5]. Considering the significance, research communities have prepared for the anticipated difficulties and required resource to move forward. And the corresponding strategies can be reflected in the research plans published as white papers [6] or formed as national-wide organizations [7, 8].

Generally, the collective behaviours of neurons generate the intelligence and bear the cognition. For certain research and applications, study from the functional level, i.e., investigation of the overall activities taking place inside the brain, is more important than focusing on the molecular or cellular level, i.e., understanding of the structure and metabolic process of a single neuron. In this circumstance, brain imaging plays the key and offers the unsubstituted assistance.

Neuroimaging technologies have undergone great developments during the past decades, and the update-to-date methods include Electroencephalography (EEG), Magnetic Resonance Imaging (MRI), functional Magnetic Resonance Imaging (fMRI), Magnetoencephalogram (MEG), Positron Emission Tomography (PET) and Transcranial Magnetic Stimulation (TMS), etc. [9]. Different technologies have different advantages and disadvantages and tend to be superior over one another in different scenarios [10, 11].

EEG is based on the fact that the neurophysiological or high-level cognitive process of brain is fulfilled via neuronal communications, i.e., neuronal electrical signals sending to or receiving from others. EEG measures these electrical signals which bear the activities of the brain, by placing electrodes on the scalp [12, 13]. This enables EEG the most convenient approach in conducting brain research and applications. From the measurement, it can tell the characteristics of brain activities, for example, how brain activity changes in response to stimuli. The high temporal resolution characteristic of EEG also elects it as the suitable assistance in diagnosing abnormal brain activities such as epilepsy [14].

The general working principle of an EEG acquisition device is shown in Fig. 1.1. Cognitive process usually happens due to thousands of neurons firing together. The sensor (electrode) on the EEG cap can detect such spontaneous electrical activities over a period. Then the signals from the electrodes are sent to the amplifier for amplification.

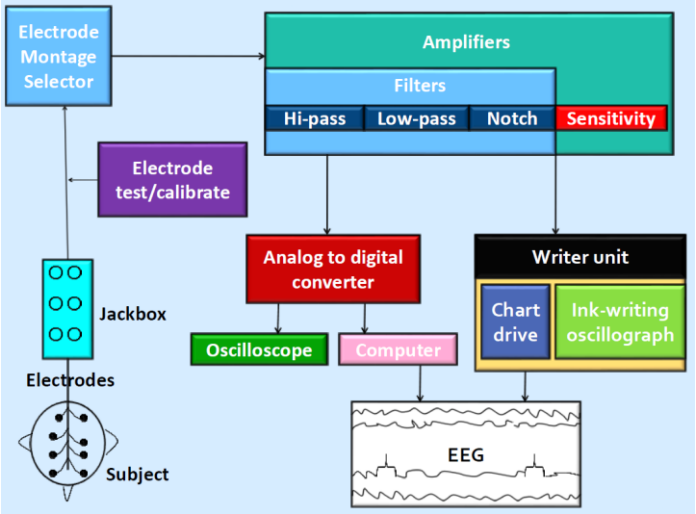


Fig. 1.1 Schematic diagram of EEG device. (Courtesy: Vajarala Ashikh)

In general, there is an auxiliary circuit to sample and digitalize the signals before transmitting to a computer. With dedicated software run on the computer, it can generate various maps of brain activities with a fine-granular temporal resolution in a rapid way [15].

For the sake of concentration and brevity, the principles and complexities of other imaging technologies such fMRI and MEG are omitted. Compared with them, the advantages of EEG lie in many aspects:

- The non-invasive and passive working principle guarantee its safety.
- EEG directly measures the brain activities, conferring a high temporal resolution.
- A simple training can qualify the operation, instead of intensive field experience.
- The mobility feature enables the device to be capable of certain immersive tests.
- The low price makes it accessible for most researchers to experiment with it.

However, due to the reasons to be mentioned subsequently, EEG signals are among the most complex biosignals, and EEG data analysis is among the most challenging statistical or machine learning tasks. Although abundant resources are devoted, too many factors clung together limit the achievements. This affects the potential applications which relies on the high accurate interpretation of EEG data, such as BCI [16, 17].

The first difficulty to perform EEG data analysis with satisfying results is signal degeneration. Due to the way of EEG signal acquisition, electrical signals generated in certain deep buried cortices must propagate through four layers of head structure to be captured by the electrodes. There are inevitable interference and attenuation which cause signal degradation and distortion during the propagating route. Usually, the size of a cortex inside brain as well as the strength of the signal magnitude vary from one subject to another. These can cause variance of EEG signals in identical experiments. Such intra- and cross-subject variance have well explored in previous EEG research [18, 19].

The second impact comes from the noise and variations during the signal capturing process. For electrode placement, the international 10-20 system [20] is recommended. However, from the practical perspective, products of available EEG cap on market always have fixed electrode positions. With varying head shapes of individuals, the mismatch between the intended electrode placement and the actual

placement can unavoidably introduce signal perversion. And the conductivity between scalp and electrodes can also impact the quality of the signal. EEG experiments tend to be endurance ones, measured by dozens of minutes to hours or even to days. The impedance of the electrode between contact interface might endure fluctuation especially when the gel is exposed to air for a longer time. However, it is difficult to decide the change of signal amplitude is due to impedance alters or brain activity varies.

The third side effect comes from the sensitivity of EEG experiments. It is known that the captured signal is the accumulation of several signal sources during brain's neurophysiological process. And eye blinks, eye movements or body movements can sponsor certain signal source and consequently introduce artefacts [21, 22]. The problem is that it is generally quite hard to keep the test subjects stay still during the long-lasting EEG experiments, especially the experiment requires subjects' certain activities. In addition, EEG device is usually power supplied by main electricity via adaptors converting from 110/220 volts to 5/12 volts. However, the exceeding 126 dB of 220-volt main power supply to 100-microvolt electrodes incur the vulnerability of signal recording process for voltage fluctuation especially voltage burst.

The fourth factor attributes to the domain knowledge itself. Data analysis usually requires domain knowledge to evaluate the viability of the model or algorithm. For instance, imaginary movement classification is a stereotypical experiment regarding BCI applications [23, 24]. However, due to the limited knowledge of the underlying cognitive process, it is still not very clear which are the novel features buried in the EEG data that can be extracted for analysis. Therefore, on one hand it relies on pre-defined novel features to boost model performance for understanding the cognitive process; on the other hand, it is required to understand the process in order to decide what could be the novel features. Such a chicken-and-egg problem does add the difficulties for EEG data analysis.

The last challenge is the so-called labelling problem. Usually the mind states of cognitive processes, such as vigilant or drowsy, are not self-evident for observation. However, to evaluate the analyzing results, some ground truth labels must be provided. Because it is hard or impossible to directly measure mind states, indirect ways are taken alternatively. But the mapping between indirect indicator and intended labels are potentially non-linear, and a rough estimation probably introduce bias into the articulated labels.

To sum it up, facing the challenges rooted in the EEG signal traits, to investigate the current DL methodologies and tools which are adept in these regards is a meaningful and useful work. It is believed that the systematic study of DNN which demonstrates remarkable achievements in many other fields, can accomplish certain EEG data analysis tasks for the current research; and the accumulate knowledge and experience can benefit biosignal processing in a long run.

1.2 Deep Learning

DL is the most highly topical research currently with outstanding achievements in various fields [25, 26]. It has the deep root in neuroscience and tends to draw inspirations from the way brain computes. It is believed that in millions of years evolution and natural selection, brain works out the methods which are optimal for solving variational nonlinear problems. By combining the biological neural facts into

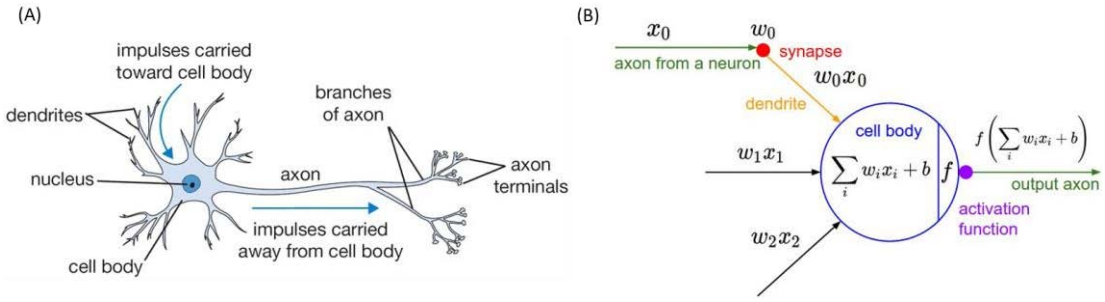


Fig. 1.2 (A) A diagram drawing of a biological neuron; (B) The corresponding mathematical model. (Courtesy: Andrej Karpathy, Stanford's CS231n lecture)

artificial neural networks (ANN), it is expected that a proximate capability could be achieved [27].

The origination of ANN can be traced back to the work done by McCulloch and Pitts who were simulating the way of neuron and neuronal systems to construct a resembling computational model. They introduced the concepts such as input, hidden and output layers, which are still used as most contemporary neural networks now. They also successfully demonstrated that calculations could be performed by a network of artificial binary neurons [28]. The limitation of their work is that weights are fixed instead of trainable. Later, Frank Rosenblatt published his work on perceptron which accompanied the first major neural computing related research project, enabled the computing model in a quite similar modern form [29]. An analogy of these preliminary research is illustrated in Fig. 1.2. The structure and working principles of the neuron in Fig. 1.2(A) is modelled as a mathematical transform in Fig. 1.2(B).

However, the route leading to an efficient method in training the neural network lagged behind the model itself. Although Donald O. Hebb presented a learning method which could adjust the weights of networks as early as in 1940s [30], it turned out to be too simple especially for multi-layer perceptron (MLP). The surge had to be waited until 1980s when Rumelhart, Hinton and Williams jointly proposed the back-propagation algorithm [30], which installed the engine for driving neural network to converge to its optimal state efficiently. Although the computational facilities severely constrained the training of large-scale neural networks in practice, some works have already paved the foundation and simply wait for the maturity of silicon to ring up the curtain. For example, Sepp Hochreiter and Jürgen Schmidhuber together invented long short-term memory (LSTM) [31]; and Yann LeCun, et al., developed the original convolutional neural network (CNN) which is the prototype for most subsequent networks [32].

With the entering into the new millennium, especially from 2008, the modern semi-conductor breakthroughs propel the speed and capacity of numeric computation into a new era. As a precursor, Hinton and his students succeeded in utilizing CNN for image classification and achieved the best result. It lit the spark of DNN, which later sweeps numerous fields promptly with records set never seen before, such as image classification [33], speech recognition [34], language understanding [35], drug discovery [36], particle accelerator data analysis [37], cancer prediction [38], exoplanets identification [39], etc.

Compared with MLP, DNN can also be regarded as brain-inspired computation. The development of neuroscience reveals this correspondence between visual system and DL. By virtue of the work from David Hubel and Torsten Wiesel [40, 41], it is now understood that the visual system processes perceptions in an iterative, hierarchical and

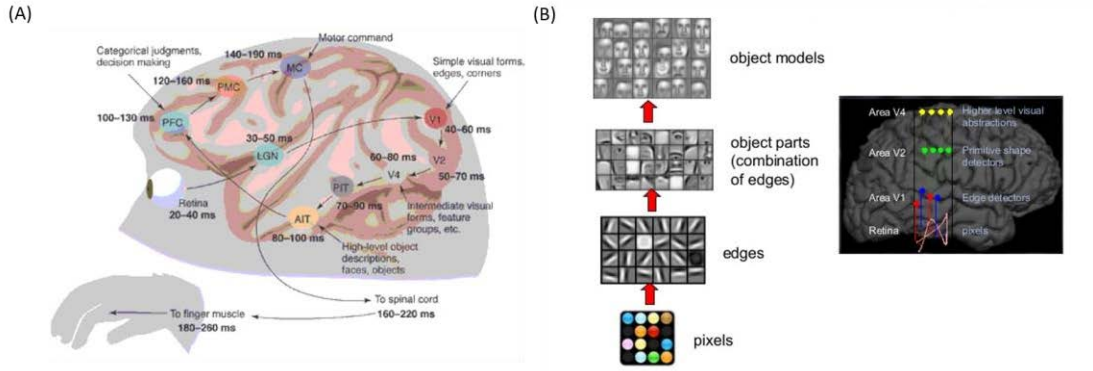


Fig. 1.3 (A) Hierarchical transmission of information from primary cortical to beginner to high-level visual cortex in human brain (Courtesy: Simon J. Thorpe and Michele Fabre-Thorpe); (B) Feature hierarchy in deep neural network (Courtesy: Honglak Lee).

concrete-to-abstract procedure, as in Fig. 1.3 (A). For DNN, via successive convolutional layers, the intrinsic features to the data itself can be effectively captured at various levels as in Fig. 1.3 (B). It is regarded that by analogy to the visual system, DNN can effectively extract features hierarchically from the raw data. Such an automatic process can save effort on the design of the feature extraction algorithms manually. Because for EEG data analysis, limited understanding of brain makes it relatively hard to decide the suitable features, DNN offers a great alternative to cater for such a dilemma.

The ongoing research of DL also incorporates or is combined with other ideas and methods for new applications, such as generative adversarial network (GAN) [42], deep Q-network (DQN) [43], etc. These works not only generate astonishing results which are thought to be some decades beyond, but also inspire works in other disciplines. For example, the idea of GAN (two neural networks, i.e., the generative network and discriminative network, battle in a game to reach the equilibrium) has been introduced into domain adaptation with promising results achieved [44]. The accomplishment of domain adaptation via DNN indicates its potential in solving intra-subject and cross-subject variance, which is ubiquitous in EEG signal processing especially from the BCI perspective. Thus, based on these achievements and considering the diverse paradigms of BCI applications, it is appealing and worthwhile to investigate these actively developing methods to analyze EEG data, and a systematic research into the methodologies of DL with networks specially tailored for EEG data can advance the EEG-based research over various perspectives.

2. Research Overview

2.1 Literature Review

There are several purposes to analyze EEG data. First, it can aid brain research or clinical diagnosis as a brain imaging technique. By studying the traits of activations, it can unveil the involvement of a specific brain cortex in a physiological or cognitive process, thus deepening the understanding of certain phenomena [45-47]. Second, it can help to identify the characteristic of brain state and the voluntary or involuntary inclination to target some BCI applications [48-50], which is believed to be the next frontier of technologies. Fulfillments of these purposes generally require the basic domain knowledge to carry on research in this regard. [4] is a renowned textbook systematically covering every aspect of neuroscience and laid the physiological fundamentals. [51] is another widely-used textbook mainly focusing on EEG signal processing. It begins from the generating mechanism of EEG signal, such as the biological basis, mathematical modelling, to techniques derived from conventional signal processing perspective for processing EEG data.

From the establishment of EEG as an auxiliary in brain study or BCI applications, numerous methods have been proposed to achieve the above purposes. Depending on these methods and algorithms, EEG data can be processed in the time domain, frequency domain or spatial domain. Fig. 2.1 illustrates the relations between different representations or formats of EEG data. It is pointed out that the transforms among different domains can be irreversible.

The most natural way to process EEG data is to take the multi-channel waveform in the time domain, and this is also the way used by most EEG acquisition devices for storing raw data. However, besides some special cases, it is impracticable to glean useful information directly from the waveform data. In practice, technicians with trained eyes from clinics, might be able to identify abnormalities directly from EEG waveform signals [14]. There exists BCI-based research that analyzes EEG data in the time domain. In this thread, one pioneer work is about the theory and practice of common spatial pattern (CSP) filters. It separates a signal into subcomponents with maximum differences in variance via simultaneous diagonalization of two matrices [52, 53]. It has achieved various well-recognized results, especially through subsequent works which introduced more tricks and techniques [54, 55]. CSP also inspires the design of EEGNet which is contemporary DNN based [56]. It is mentioned that one

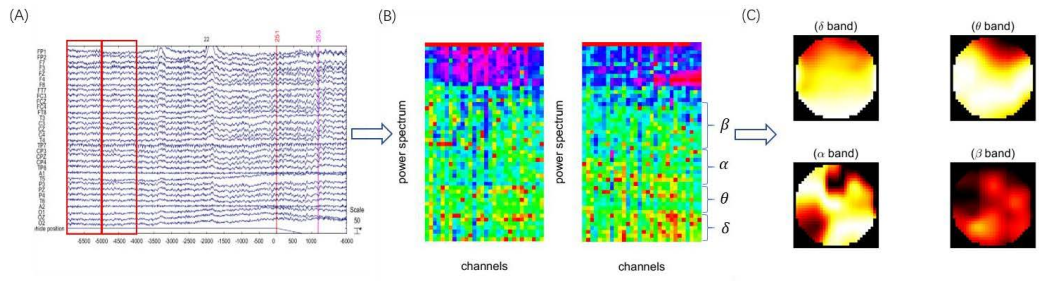


Fig. 2.1 EEG data format (A) waveform data in time domain (B) power spectrum in frequency domain (transformation of segmented waveform data) (C) EEG topograhpies in spatial domain (mapped from power values of same subbands of multiple channels)

advantage for directly processing waveform data is the avoidance of computation overhead enforced by domain transform if processed in other domains.

Research-oriented EEG analysis tends to be carried out in the frequency domain. EEG signal oscillations along the time axis are generally the reflections of underlying brain activities, and noises caused by artefacts or other factors are unavoidable. By scarifying the temporal information, certain features such as frequency distribution and power spectrum are more eminent than wave bursts or oscillations in the time domain. Because transforming from the time domain to the frequency domain requires EEG data lasting for some duration for processing, it means some unusual variance is discarded at the same time. Based on these more stable spectrums, more sophisticated biomarkers can be achieved [57, 58]. For instance, related treatments such as the event-related potential (ERP) and the event-related spectral perturbation (ERSP) have underpinned many physiological and cognitive discoveries [59, 60]. Notably, by introducing probability theories, above methods can fit into broader frameworks, either categorized as parametric methods or non-parametric methods. Transform from the time domain to the frequency domain is usually based on discrete Fourier transform (DFT) or fast Fourier transform (FFT), while wavelets transform is another option. But wavelets analysis tends to jointly analyze in both the time and frequency scale, might not be a pure frequency-oriented method [61]. In practice, specific frequency bands, such as delta, theta, alpha, beta respectively might have special interests for the research community to correlate to specific brain activities or functionalities.

For EEG data acquisition, different EEG montages (including references) can be adopted for recoding signals [62]. Therefore, EEG signals captured from similar processes can present quite different waveforms. By mapping brain activities into topographical distribution, it can avoid processing the unimportant electrically neutral reference point. EEG topography can display a vivid illustration of EEG power distribution over the scalp. In the circumstance where brain function and corresponding connections with different vortices are studied, working with EEG topographical data is the most convenient way [63, 64]. In addition, EEG topography has important clinical implication. By studying the spatial distribution of frequency spectra jointly with other methods, the practitioner can qualitatively or even quantitively study the situations of patients, especially from the pathological perspective [65, 66].

It is pointed out that in order to suppress the artefacts and increase the signal-noise ratio (SNR), pre-processing of EEG signals is generally required before further processing EEG data. Fig. 2.2 illustrates the conventionally conducted pre-processing

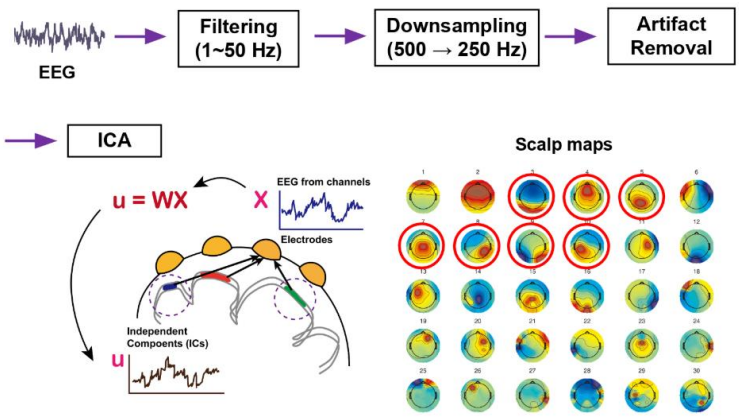


Fig. 2.2 EEG data pre-processing. (Courtesy: Dr. Yu-kai Wang)

procedures. Besides that, normalization might also be applied to mitigate data variance. Notably, the pre-processing might be regarded as the preceding steps of feature engineering via a hand-crafted way. Without a thorough understanding of the working principles of brain, there might be bias in doing so. For example, in Fig. 2.2, the selection of the ICA components is mainly based on previous research or empirical experience [67]. It might impact the conclusion if there is bias in the previous understanding. Another concern is about normalization, it could only be done off-line for some applications. Generalizing the model into online application might be problematic due to inapplicable normalizations.

Therefore, by studying the achievements of pioneering work as well as the limitations, it is planned to systematically investigate various aspects of DNN for EEG data analysis. It is expected by taking the auto feature extraction capability of DNN, the limited understanding of neurophysiological and cognitive process can be improved via the end-to-end learning paradigm. It is also anticipated by adopting innovative ideas in DL such as GAN, some problems such as intra-subject and cross-subject variance which is a critical challenge for conventional methods to be effective. The achievements of DL hitherto shed the optimism that DL can potentially boost BCI applications with promising results.

2.2 Challenge Highlights

Even EEG data is not profoundly different from the datasets where exhilarating results are achieved via DL, the corresponding challenges need to be emphasized as criteria in DNN design and metric for performance.

The first challenge concerns problem formulation. For commonly encountered scenarios of DL, such as image classification or object detection, they are straightforward from the application perspective. For BCI, experiments need to be studied to convert terminologies in the neuroscience or BCI to the language of machine learning (ML). During the study, it requires to figure out the link between the goal of experiment from the EEG perspective, and the fulfillment from the ML perspective. Usually, the experiments need to be carefully checked to make sure the theoretical postulation, practicability constrains, output predictability, etc., have full correspondence in the ML or specially, DL context. Several stereotypical experiments such as P300 speller experiment [68], oddball experiment [69], driving fatigue experiment [70], mind-load experiment [71], etc., originate and manifest themselves in understanding certain aspects of neuroscience via EEG. It requires to interpret these experiments to shape them into ML problems. Another reason is that the understanding of the basics of EEG experiments can aid the diagnosis of network models, interpretation of the results and prediction of potential discoveries.

The second challenge is that the coincidence between tasks of EEG data analysis and tasks of DL. It is not as obvious as conventional image processing tasks and ML tasks. The conventional image processing relies on quality filter design, which can be accomplished automatically by DL. However, for EEG data analysis, various ways such as matrix decompositions [52], transformations and integrations [72], statistics [73], etc., can be utilized, and it is hard to reach a correspondence from DNN viewpoint. Although there are works done in this regard [74-77], the current applications of DNN for EEG data analysis are still limited. There lack investigations in two aspects: (1) The unification of EEG data analyzing tasks and the DNN modelling methodologies. (2) Different network structures and their effectiveness for EEG data analysis. For example,

it is estimated that the recurrent structure can capture up to 20 words of a sentence [78], but it is not clear this capacity is sufficient for handling complex EEG data.

The third challenge is that for EEG experiment, samples are costly to obtain. It needs the high cooperation from test subjects recruited under ethical approval, which sometimes is time and resource consuming. For DNN, it depends on the large number of samples to drive the convergence of network during training. However, with limited data samples, consecutive runs with the same samples increase the potentiality of overfitting. It is understandable that with such large quantity of DNN parameters, limited samples incur the inclination that the network learns to remember the data instead of learning some common representations. It has negative impact on generalization and the case can be severer for EEG data.

Furthermore, the general sample augmentation methods facilitated in image processing can be inappropriate for EEG data. Usually, to increase the generalization of DNN, there can be data augmentation to the original samples such as noise-adding, rotation and flipping, etc [79]. However, it makes no sense especially for EEG topographical data since it has strict neurophysiological meanings. The augmented data is of great chance biased. For example, in Fig. 2.3, the flipping of a dog image can still be a valid input; however, a flip of EEG topography image might never appear in real world, even with the same experiment scenario by recruiting people with opposite handedness.

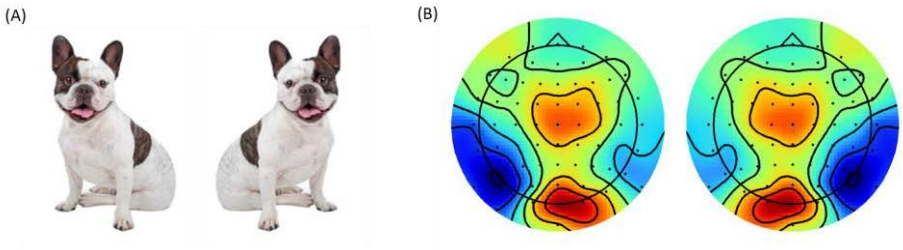


Fig. 2.3 (A) Flip of dog image (valid); (B) Flip of EEG topography (severely biased).

The fourth challenge is that the distribution of training set can be quite different from the test set for EEG data, contrasting with general tasks where training set and test set coincides, a key for satisfying DNN performance. Generalization requires that the test set is obtained by sampling from the same distribution as the training set [80]. However, the manifest intra-subject and cross-subject variance of EEG signals are failed in this respect, i.e., these variances incur the distribution discrepancy between training data and test data. It might post the difficulty in diagnosing the performance of designed network architecture. If the performance goes as unexpected, it is difficult to diagnose the root cause, for example, the architecture or the data attributes to the failure.

Finally, it is expected the analyzing results from DNN model can reciprocally aid the understanding of the corresponding cognitive process, such as the activation of specific brain cortex during the experiment. However, the lack of theoretic conclusions about DNN, such as the interpretability as well as neural correlate, its dynamics and the black-box operations present difficulties in interpreting the learned filters and feature maps [81]. It still needs to investigate and find out efficient approaches in this regard.

To sum it up, the success of DNN in various applications inspire the application for EEG data analysis, although challenges accompany the whole process, such as problem forming, modelling and performance assessment. But it should not undermine the study of DL with its application to different aspects of EEG data analysis. Besides,

new methodologies such as GAN and DRL indicate the potential of tackling the difficulties in EEG data analysis which are ineffective for conventional approaches, so it is optimistic to expect promising achievements via research in this regard.

2.3 Stereotypical BCI Experiments

Brain imaging via EEG is an indispensable element in BCI applications. By studying EEG data captured during well-designed or well-controlled BCI experiments, it can draw insightful conclusions towards cognitive conjectures or assumptions, and knowledge gained from this can facilitate the smooth and effective interaction between brain and computer or machine. To quantitatively and qualitatively benchmark the performance of proposed DNN architectures for EEG data analysis, three stereotypical experiments are introduced in the following subsections. They are frequently referenced to form the problem, to understand the data, to test the model and to interpret the result in ML context. Using common experiments enable the chance to objectively measure the suitability and efficiency of the model under different situations. These experiments are conducted by us or our collaborators participating in the collaborative technology alliance (CTA) on brain research, or by third-party scholars.

2.3.1 Driving Fatigue Experiment

Fatigue is regarded as the most severe factor causing road fatalities [82], and one manifestation of fatigue during driving is drowsiness. One promising research direction of driving safety is by monitoring drivers' mind state via portable EEG device [48, 70]. Based on the study of correlations between fatigue (especially drowsiness) and driving performance revealed by the captured EEG signals, a portable monitoring device could be developed to alarm on fatigue driving.

To target a practical application, a simulated endurance driving experiment is conducted in our lab to firstly simulate the problem, explore potential solutions, and assess the proposed method. It is designed as recruited certified drivers wearing EEG cap, performing an approximate 90-minute driving along a virtual four-lane highway. It is postulated that during the experiment, subject's mind state, such as alertness or drowsiness, is hardly maintained all the same. Fig. 2.4 illustrates the experiment setup. To simulate different driving conditions on the road, a real car is converted and installed on a manoeuvrable platform with six-degree freedom to guarantee sufficient flexibility, as in Fig. 2.4(A). The horizon for the driver is projected on a large chained screen to mimic the highway scene as in Fig. 2.4(B). The experiment was approved by the Institutional Review Board of the Taipei Veterans General Hospital (PN: 101W963, VGHIRB No.: 2012-08-019BCY), and the voluntary, fully informed consent of the subjects participated in this research was obtained as required the by accompanying regulations handled for ethical committee review (IRB -TPEVGH SOP 05).

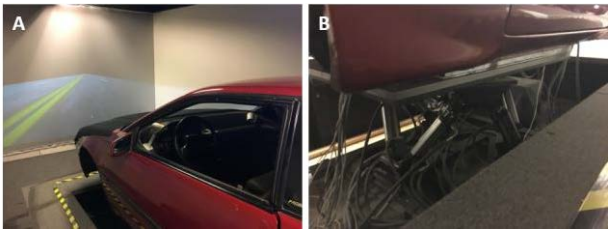


Fig. 2.4 Simulated driving setup: (A) the mimicking highway scenario; (B) the maneuvering platform with installed converted car.

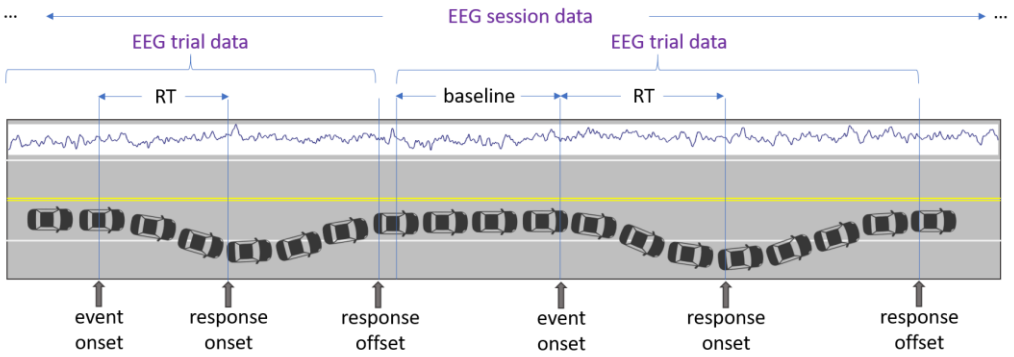


Fig. 2.5 Illustration of the experiment terminologies.

The procedures of the experiment are as follows. On usual the subjects operate the car cruising along one lane. Then a deviation to the car is deliberately introduced (event/deviation onset) to have the car randomly drift leftward or rightward. The subjects are instructed to move the car back to the original lane by steering. The moment of subject's acting is called response onset. The duration lasting from event onset to response onset is called reaction time (RT), treated as an indicator for subject's drowsiness. It is postulated that longer RT corresponds to deeper drowsiness. Once the car backs to the original cruising state, it is regarded as response offset. To correlate drowsiness with driving performance, the driver only needs to operate the steering wheel and is freed from brake and petrol panels controlling.

The traversal from deviation onset, response onset to response offset is called a trial. New trial begins about 5~10 seconds at random after the previous trial. The whole session, i.e., subject's once participation in the experiment, can contains many trials. The corresponding terminologies which are key to understanding the experiment are illustrated in Fig. 2.5. Notably, the number of trials in one session is highly depending on the situation when subjects participated in the experiment. For example, if the subject mostly experiences sleepiness during the experiment, the bulk of trials could last longer due to poor reaction. Because the whole process is always around 90 minutes, varieties of RT can induce different number of trials among separate sessions.

EEG data was captured simultaneously and continuously during the whole process using Scan SynAmps2 Express system. For specification, EEG cap is with 32 Ag/AgCl electrodes, of which 2 electrodes are for reference (opposite lateral mastoids). The placement of EEG electrodes is according to a modified international 10-20 system. The contact impedance between all electrodes and the skin was kept <5kΩ, with sampling rate 500 Hz.

2.3.2 Visual Oddball Task

Visual evoked potential (VEP) is an important neurophysiological terminology which has both clinical significance and BCI-oriented importance [83]. One well understood phenomenon related to it is P300, an event-related potential (ERP) coupled to subjects' exposure to expected visual stimuli. It is the cornerstone for BCI applications such as P300 speller, which is by randomly flashing on and off letters and analyzing the corresponding EEG wave, helping subjects or patients directly communicate with outside world via brain [68].

The visual oddball experiment is conducted by our collaborators [84]. The experiments were approved by the U.S. Army Research Laboratory (ARL) Institutional

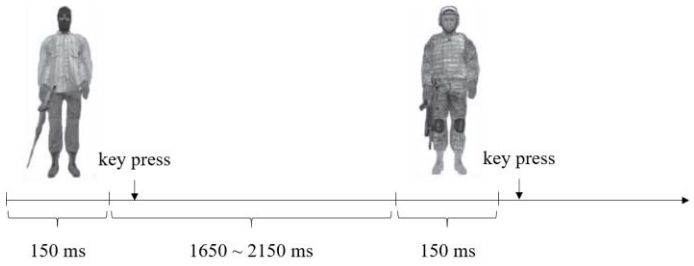


Fig. 2.6 Experiment paradigm of visual oddball task

Review Board (Protocol # 20098-10027) with consent of subjects complying with the Federal and Army Regulations. The task is designed as image stimuli, which are of two categories (target vs non-target) presented to subjects at a rate of 0.5±0.1 Hz (approximately one image every two seconds) to arouse the responses. For this experiment, the targets are enemy combatants and non-targets are U.S. soldiers. The target set is with total number 34, and non-target set is 236. During the experiment, the subjects were presented with all 270 images. They were instructed to identify each image as being a target or not with a unique button press as quickly, but as accurately, as possible. The paradigm of the experiment is shown in Fig. 2.6.

There were eighteen participants in the experiments, which lasted on average 15 min. Experiments were repeated with signals recorded by different EEG devices. Unless explicitly specified, it is always assumed that the signals captured with the wired 64-channel ActiveTwo3 system (sample rate set to 512 Hz) from BioSemi are used for analysis. Due to corruption or poor responses, data from four subjects are exempted. For pre-processing in time domain, the signals are first down-sampled to 64 Hz, following segmentation into [0, 0.7] second interval time locked to the stimulus onset. To mitigate the effect of outliers, epochs with incorrect button press are removed. Then the mean baseline is removed from each channel in each epoch. These procedures lead to a final dataset consisting of 382 target trials and 3249 non-target trials to be used in the subsequent methodology section.

2.3.3 Working Memory Experiment

Working memory is an indispensable cognitive ability which aid people's performance during complex tasks such as reasoning, comprehension and learning. It refers to the mechanism of the brain in holding information temporarily especially for immediate need [85]. To investigate aspects such as the capacity, influential factors, etc., can lead to better understanding of the working memory to help improve the cognitive performance.

The working memory experiment was conducted by third party scholars [71]. It is to measure mind load or cognitive capacity related to working memory. The experiment is to have test subjects decide whether a randomly showing letter is belonging to the previously displayed randomly-generated letter set or not. For fixed memorizing period and retaining period, the mind load is regulated by the size of the letter set, or the number of letters contained in the set. The postulate of the experiment is that brain activities could differ under different cognitive loads. A precaution for low performance of test subjects' participation is by only counting the correct experiment trials.

The paradigm of the experiment is shown in Fig. 2.7. A display of random letter set launches the new trial, and the set retains for 500 milli-seconds (ms) to consolidate

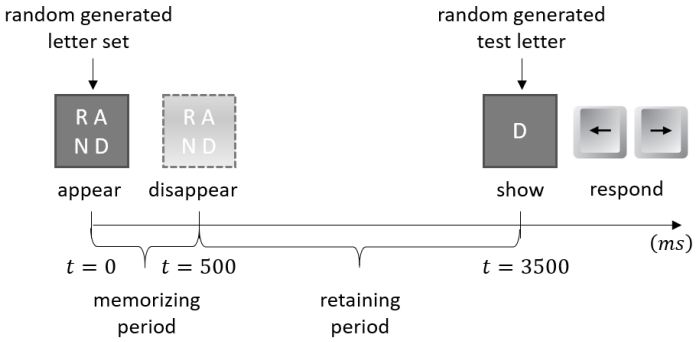


Fig. 2.7 Working memory experiment paradigm.

the subjects' memory. After 3 seconds(s), the subjects are given the chance to decide if a randomly popped up letter belongs to the previous letter set or not, by hitting specific key. New trial will begin after random delays. There were 13 subjects who participated in the experiment.

EEG is recorded along the process simultaneously and continuously using Neuroscan via 64 sintered Ag/AgCl electrodes placed with accordance to the 10-10 montage, with reference to an electrode placed ~ 1 cm posterior to Cz. An online filter is applied to down-sample the original signals of 500 Hz to 250 Hz. Electrode impedance was maintained $\leqslant$ 5 kΩ over the during the experiment. The experiment and data capturing specification can also refer to [86].

2.4 Research Map

Anticipating the results to be achieved by utilizing DNN and its generalizations for EEG data analysis, this research plans the DL centred methodology and goals as reflected by the research map in Fig. 2.8. The map consists of four regions, each represents the respective methods and goals. The circled numeral landmarks correspond to target works in the following chapters to fulfill the plan. Overview of the methods and goals in each part of the map are listed as below and are also referenced to in following chapters:

① Method: By reviewing the commonly-used network structures and considering the characteristics of EEG, building a typical DNN as in the general DL approach to analyze EEG data.
Goal: To benchmark the performance with other methods while to seek the interpretation of learned filters.

② Method: By formulating the EEG data variance problem in the context of domain adaptation and investigating the idea and practice of GAN, tailoring a special DNN for adversary realization.
Goal: To reduce the intra-subject and cross-subject variance of EEG data.

③ Method: Studying memory networks with the consideration from EEG data traits. Interpreting DeepMind work, i.e., DNC, to propose a stacked version and employing the self-organizing maps (SOM) as the memory module.
Goal: To either effectively harvest EEG information buried in the endurance experiment and to boost the performance. To take advantage of limited EEG samples and to relieve operational overhead.

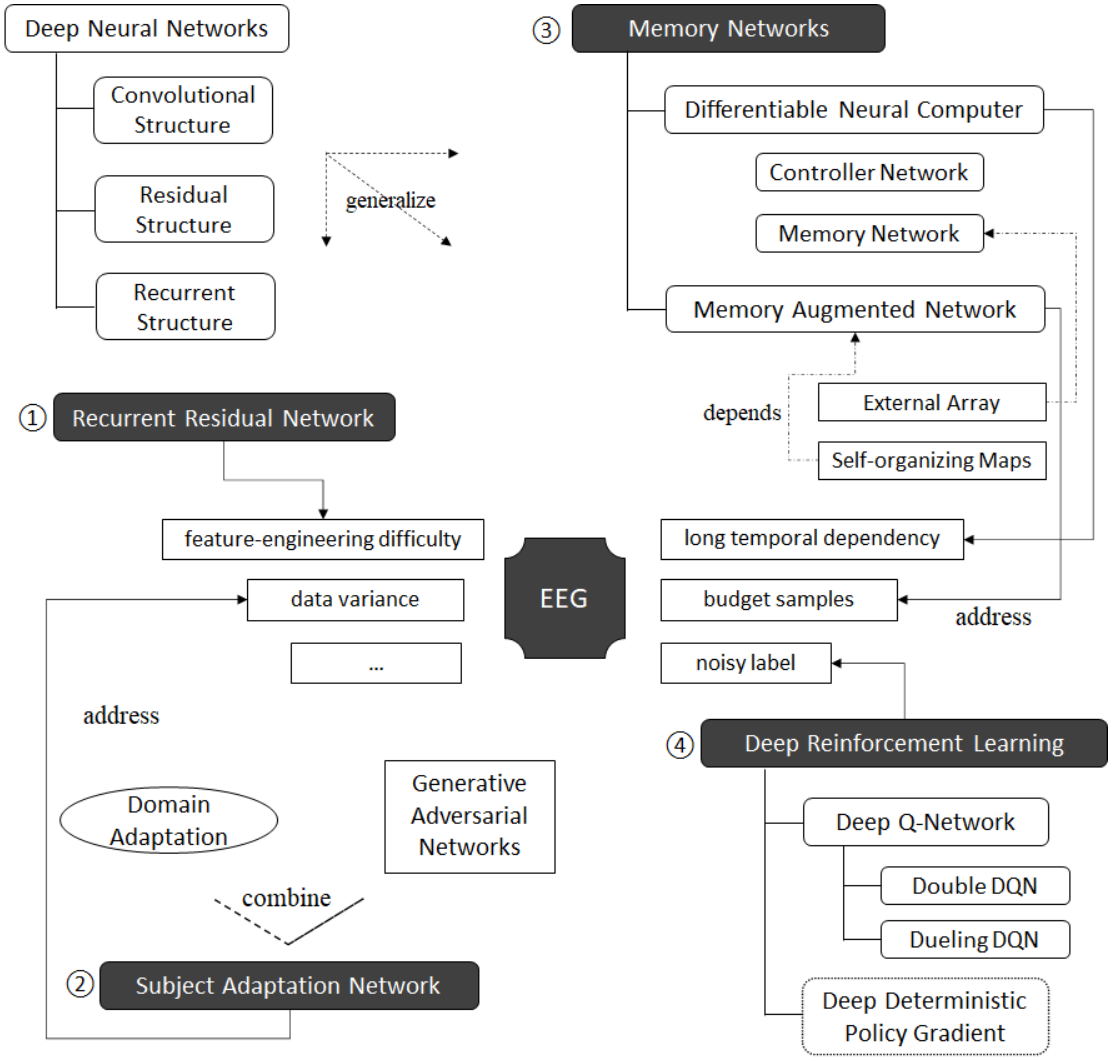
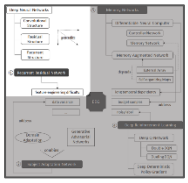


Fig. 2.8 Research map of generalizing DNN for EEG data analysis.

④ Method: By investigating deep reinforcement learning and its tolerance to labels, introducing a new computation paradigm exemplified by DQN for certain BCI applications.
Goal: To mitigate the noisy label problem and to trace the procedural change of EEG data in a global and satisfying way.

To fulfill the setting goals, following chapters research into different aspects of DL to address the divergent challenges of EEG data analysis, as guided by the research map. It is expected that this systematic study of the applicability and potentially of DL applications in EEG data analysis domain not only convey a fruitful doctorial work, but also benefit subsequent research in these threads.

3. General DNN Approach



This chapter covers the top-left region of the research map in Fig. 2.8. By considering the contemporary DL theories and practice in solving specific domain problems, this chapter investigates and studies the applicability, performance, and potential interpretation of DNN for EEG data analysis in a general approach.

3.1 Basic Structures Review

To propose neural network structures suitable for EEG data analysis, this research begins from the systematic investigations of basic network structures for their specialties in matching the characteristics of EEG data.

An indispensable component in DNNs is the convolutional structure. Convolution plays an important role in the conventional signal processing, where filters of different properties are designed to manipulate the signals for specific purpose [87]. Hence, nothing can be overemphasized on the importance of convolution. EEG data analysis especially from the time domain falls into this category. Convolution in DL still undergoes progressive development. It evolves from the original vanilla convolution, to the depthwise convolution and the separable convolution [88]. The operations are illustrated in Fig. 3.1. For a specific problem, different convolutions can be varied to seek the optimal one. In general, if individual channels are more independent of each other, depthwise or separable convolutions might be a preferable choice.

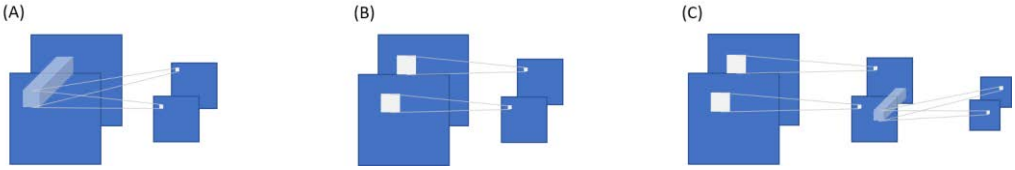


Fig. 3.1 (A) vanilla convolution (B) depthwise convolution (C) separable convolution

Recurrent structure or recurrent network is another important component which is prominent for processing time series data [71]. Recurrent structure has different originations in different disciplines. For example, the form of it in engineering is not taking the exact form in biology. In traditional signal processing field, the concise form of infinite impulse response (IIR) is in a recurrent expression. It is efficient in computation but complicated from the designing perspective [87]. For biology, the neuronal circuitries are ubiquitously endowed with recurrent structures which are of paramount importance for the physiological or cognitive functionalities. Two basic types are shown in Fig. 3.2. Recurrent structures in these disciplines inspire the recurrent neural networks (RNN) in machine learning. RNNs are used to capture the

Fig. 3.2 Biological recurrent structures (A) Inhibition (B) Excitation

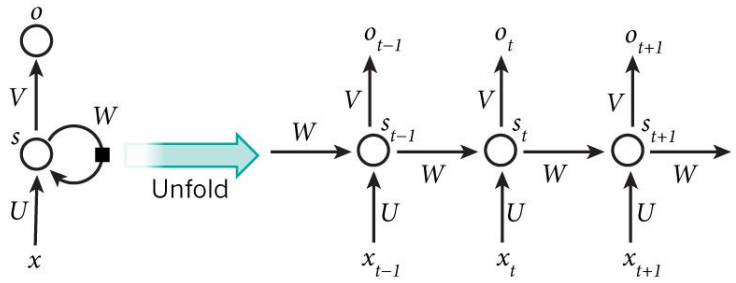


Fig. 3.3 Recurrent neural network and its equivalent unfold along time axis. (courtesy: Denny Britz)

temporal dependencies between steps/frame of time series data. An illustration of the compact structure and its unfolding along time axis is in Fig. 3.3.

The instantiation of a specific RNN depends on the problem itself, especially the relation between the sequence samples and labels. The common paradigms could be chosen from are shown in Fig. 3.4. In addition, for the same structure, the cell itself can vary from one type to another. The original vanilla cell, which is represented by a self-adjoint fully connected matrix, can induce strong coupling between two consecutive steps. And as early as in 1997, Sepp Hochreiter and Jürgen Schmidhuber introduced the long short-term memory (LSTM) to modulate the correlation between time steps in a finer manner [31]. Recent improvements are mainly from the simplification perspective to mitigate the computation overhead [89]. For example, introduction of grated recurrent unit (GRU) reduces the number of gates but still keeps the essential component of LSTM. These commonly used cells are shown in Fig. 3.5.

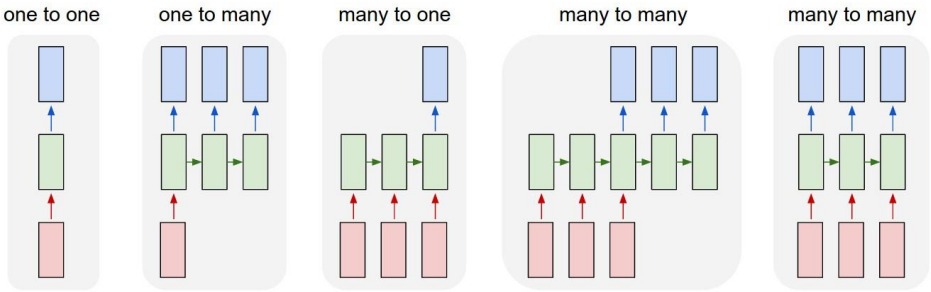


Fig. 3.4 RNN paradigms decided by the characteristics of sequence. (courtesy: Andrej Karpathy)

Besides CNN and RNN, new architectures of DNN are proposed during the research development; for instance, the residual structure in Fig. 3.6(A) [90], and dense structure in Fig. 3.6(B) [91], etc. The residual structure is motivated by the insight of learning the essence. Considering an extreme case where the identity was the optimal mapping, it would be a better choice to learn nothing instead of learning an identity by stacking of nonlinear layers. The reduction on unnecessary training enables residual structure exceptionally effective in stabilize the learning process. For the dense structure, it can be viewed as an extension to residual structure. The biological correspondence might be referenced to the work in [92]. In addition, practices reveal

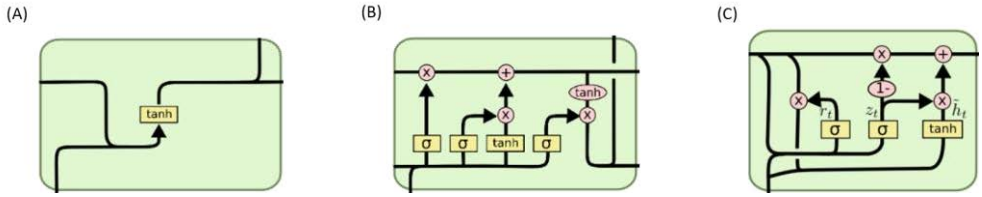


Fig. 3.5 (A) Vanilla cell; (B) Long short-term memory cell; (c) Gated cell (courtesy: Christopher Olah).

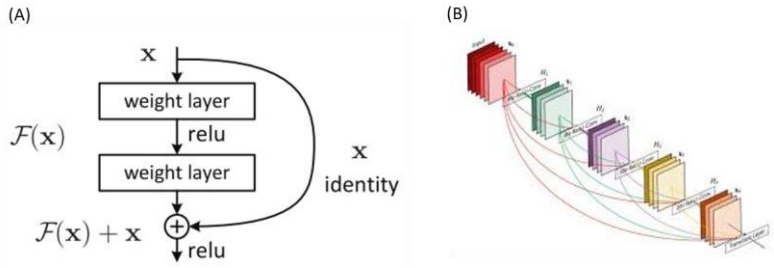


Fig. 3.6 (A) residual structure (courtesy: Kaiming He); (B) dense structure (courtesy: Gao Huang).

that dense structures can aid extracting more suitable features. To effectively assemble the above structures is critical in designing the appropriate deep neural network for EEG data analysis.

It is noted here that the residual structure indeed has a biological correspondence in brain, as shown in Fig. 3.7. Information originated from the entorhinal cortex arrives and is projected into the hippocampus through the perforant pathways via two ways. In the direct pathway, neurons in layer III of entorhinal cortex project to the distal dendrites of CA1 pyramidal neurons without intervening synapses. In the indirect tri-synaptic pathway, the information originated from layer II of entorhinal cortex arrives at the granule cells of the dentate gyrus. The granule cells modulate and relay the information to the pyramidal cells in area CA3 of the hippocampus. The CA3 cells send the information to pyramidal cells in CA1 via the Schaffer collateral pathway [93]. When compare the hippocampus circuitry in Fig. 3.7 with the residual structure in Fig. 3.6(A), the similarity between the two is manifest.

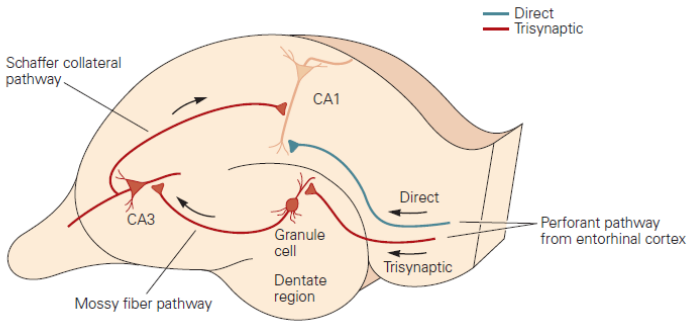


Fig. 3.7 The synaptic circuitry of the hippocampus.

3.2 Recurrent Residual Network

To propose a neural network more suitable for EEG data analysis based on the examinations of network structures above, some tactics of conventional EEG signal processing are highlighted as further inspirations. Recall that a typically adopted treatment is normalization or its variants. As shown in Fig. 3.8, a mean value for each channel in the baseline range is calculated, and values in the analyzing segment or range are offset by this mean value. In detail, for a given channel p, let $m_p = \frac{1}{|T_b|}\sum_{i\in T_b} v_p[i]$, then $\bar{v}_p[j] \doteq v_p[j] - m_p, \forall j \in T_a$. The averaging normalization could be done in the time domain or frequency domain depending on the situations. Either case, the motivation is to counter-bias the different power levels of signals among different subjects participating in the same experiment. Due to individual physiological traits or impendence caused by contact variations between electrodes and scalp, the average

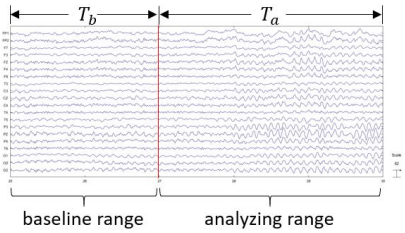


Fig. 3.8. Averaging normalization by subtracting baseline.

voltage level in the signal varies. However, it is wise to learn the essential alteration pattern of the signal, but not the direct component of the waveform data.

By reflection of the work in [90], It can be inferred that the above treatment of EEG data to some extent coincides with the residual structure as in Fig. 3.6(A), of which the motivation is to learn just the essence. For a comparable interpretation for the EEG case, note in Fig. 3.6(A), there are no weights associated with the bypass X branch, and the learned part is $\mathcal{F}(X)$. Let $\mathcal{H}(X) = \mathcal{F}(X) + X$, the equivalent learned part is $\mathcal{F}(X) = \mathcal{H}(X) - X$, or some centered or normalized version learned. The above explanation encourages the residual structure to be adopted as the fundamental building blocks for designing a neural network for EEG data analysis.

The above investigation and inspirations encourage the proposal of a recurrent residual network (RRN) for EEG data analysis in Fig. 3.9. Fig. 3.9(A) shows the whole network represented in a succinct form. The backbone of the network is the repetitions of the computing block (CB), which consists of a recurrent residual blocks (RRB) and an adjusting convolutional block (ACB). Fig. 3.9(B) depicts RRB which is the fabric of the network in detail.

For each CB in Fig 3.9(A), it is bounded by the dotted box, which recursively confines the RRB in a dashed box. Fig. 3.9(B) illustrates RRB in a time-unfolded manner. Let X_l denote the output from the bottom layer, S_l^{t-1} denotes the state of the block at a previous time step, the update of the block at layer l and time t is governed by (3.1) and (3.2):

$$S_l^t = \text{relu}\left(W_l^i \cdot X_l + W_l^s \cdot S_l^{t-1} + b_l\right) \tag{3.1}$$

$$O_l^t = S_l^t + X_l \tag{3.2}$$

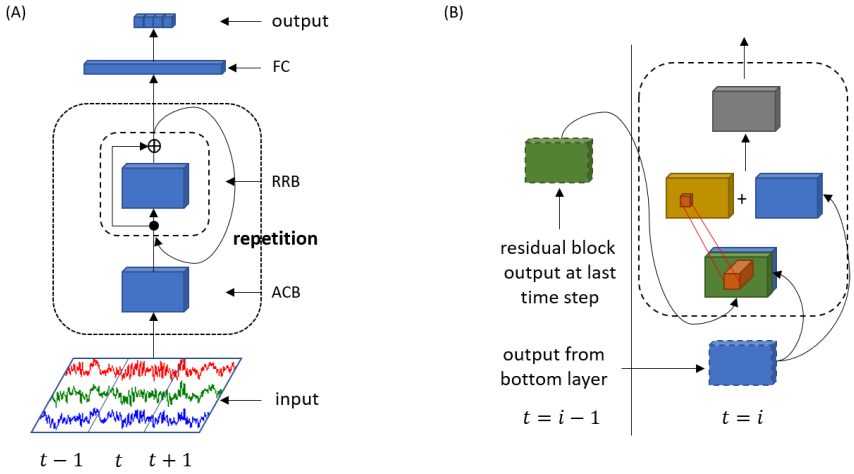


Fig 3.9 (A) The overall architecture. The dotted box defines the boundary of CB and the dashed box defines the boundary of RRB; (B) RRB with internal convolutional operation.

In practice, to explore the relationship between multiple channels, the linear transformation in (3.1) is taken place by convolution. For efficiency, X_l and S_l^{t-1} are concatenated together and applied convolution as in (3.3):

$$S_l^t = \text{relu}\big(\text{conv}\big(X_l \odot S_l^{t-1}, W_l\big) + b_l\big) \tag{3.3}$$

The inserted ACB which coupled with RRB is rooted in the residual structure. For RRB update, S_l^{t-1} and X_l must have the same dimensions to be added together. If adjacent chained RRBs are designed with different feature map dimensions and/or channel numbers, ACB, which is a standard convolutional block, must be introduced to adjust the feature map properties. For example, if it is needed to reduce the resolution of feature maps from the previous layer, this can be fulfilled by setting the stride parameter greater than 1 in ACB. Repetition of CB enhances the learning capability of the network, and the number of repetitions can be customized along with the dimension of the targeted problem.

The inputs to the network are the prepared segmented multi-channel EEG data. Each row represents one-dimensional (1D) EEG data of a given channel. However, stacking multiple channels data along the row breaks the spatial correlation between channels. Hence, all involved convolutions are 1D convolution along rows. In addition, due to the need of channel selection in certain circumstances, channels involved in the input EEG data can be fewer than the original number of channels during data acquisition. The purposes of channel selection can be attributed to the intention of investigating a specific cognitive functionality, or to remove certain channels where data are dominated by noise and becomes useless. The output of the topmost RRB is flattened and fed into two fully-connected (FC) layers. The first layer is to adjust the dimension, and the final layer caters for the classification or regression.

3.3 Experiment and Result

To justify the effectiveness of the proposed model, an EEG dataset captured from section 2.3.1, i.e. endurance driving experiment is used for evaluation. For convenience, the number of sessions for subject i is denoted by $N_S^{(i)}$. And the number of trials accumulated for all $N_S^{(i)}$ sessions is denoted by $N_T^{(i)}$. Generally, EEG data undergoes certain pre-processing, which is describe in section 2.1, for later analysis by various models. In this part, only very limited pre-processing is applied. In detail, data of corrupted channel are firstly removed from the original captured data based on manually channel selection. Then it is down-sampled to 250 Hz, followed by a band-pass FIR filter with 0.5 to 50 Hz. At last, EEG data in specified duration (6 seconds in this work) in the baseline segment immediately before the event onset are extracted into epochs for analysis.

After pre-processing, epoch data or samples from different subjects are blended and then divided according to the ratio 0.8:0.1:0.1 for training, validation, and testing. Considering that the samples between different subjects are imbalance and to prevent a biased learned model, criteria such that $N_S^{(i)} \geq 2$ and $N_T^{(i)} \geq 300$ are introduced. Based on the criteria, samples from 7 subjects are selected for analysis. It results in that 2072 samples are split for training, 259 samples for validation and testing respectively. 10-fold validation is performed to ensure the objectiveness of the assessment.

For the label preparation, RTs are mapped into different fatigue levels (represented by drowsiness) according to certain thresholds. As explained in the section

2.3.1, the experimental paradigm ensures the primary factor impacting driving performance is drowsiness and additional factors can be safely neglected. In this part, a binary classification problem is considered, i.e., to distinguish between normal and drowsy state. Hence, RT greater than 2.1 seconds is labelled as fatigue, while RT less than 0.7 seconds is labelled as alert. Labels are expressively denoted by Y to facilitate descriptions.

For the configuration of proposed RRN, all filters for ACB have the same length of 3, and their input channel number and output channel number are determined by the adjacent RRBs. All the filters for RRB have the same length of $l = 21$, and their channel numbers begin from 8 and doubled every two CB. The repetition of CB is 12, hence the numbers of feature maps are $8 * 2^{\left\lfloor \frac{n}{2} \right\rfloor}$ with $n = 0, 1, \cdots 11$. The 6-second epoch data extracted from baseline range as in Fig. 2.5, are divided into six non-overlapping segments with each lasting 1 second and consecutively input into RRN. Notably, the RRN in this case is a many-to-one paradigm, with the span covering 6 time steps when unfolded.

For benchmarking, one conventional machine learning method, i.e., support vector machine (SVM), is used for comparison. Traditional approaches generally rely on intensive pre-processing of EEG data to draw independent features for later analysis. They might be impractical for some applications. In this part, one purpose is to justify the feasibility of directly utilizing the waveform EEG data, hence, the inputs to SVM are the flatten version of the epoch data. About the configuration of SVM, the constraint is relaxed to the most margin violations with parameter $C = 1$.

For DL models, multilayer perceptron (MLP) and convolutional neural network (CNN) are adopted as benchmarks. These two networks are also capable of automatic feature extraction; however, the capabilities which closely depending on the structures may vary. For MLP, it consists of five fully-connected layers. The number of nodes in each layer begins from 4096 and decreases by half, and inductively, reaches 256 for the topmost hidden layer. The inputs into the model are the overall 6 seconds channel-wise EEG data. For CNN, a diminished version of the proposed RRN is used. The recurrent structure is restricted to 1 step, which is equivalent to a standard convolutional structure, while meantime all other configurations are retained. To cater to this change, the input segments are stacked together to feed into the network all at once.

With the data preparation and model configurations above, all models are trained with the batch number 256 for 1000 iterations, which is approximately

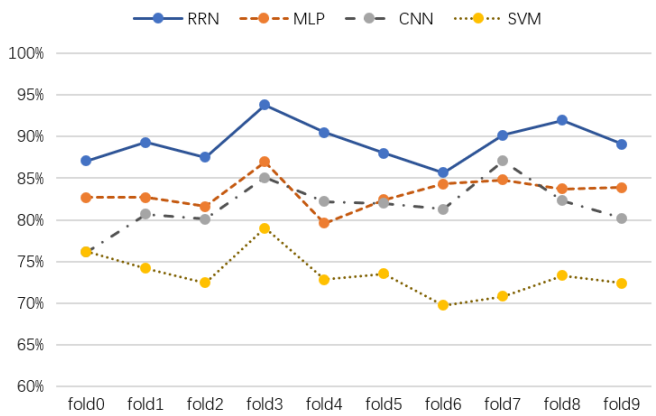


Fig 3.10. Test accuracies of different models for each fold.

equivalent to 125 epochs. Then testing samples are used for evaluating the performance of individual models. Fig. 3.10 illustrates the experimental results.

It can be observed from Fig. 3.10 that the test accuracy of the proposed RRN is supreme over all other models for each fold. As indicated, SVM relying on manual-crafted features has difficulty in achieving satisfying result. For MLP and CNN, their feature extraction capabilities assist their performance to surpass SVM. However, when compared with the proposed RRN, which is endowed with all the characteristics of MLP and CNN, the performance is undoubtedly the best among all, as shown in Table 3.1.

Table 3.1 Average test accuracies of different models

Model	SVM	MLP*	CNN	RRN
Accuracy	0.734	0.832	0.817	**0.893**
Improvement	-11.7%	0.0%	-1.8%	**7.3%**

*base unit to calculate relative improvement by comparison

3.4 Discussion and Future Work

It is worthwhile to investigate the potential correspondence between the learned filters and filters used in the conventional EEG data analysis. Convolutions (and the associated filters) can be interpreted as local template matching in time domain or spatial domain. In the spatial domain, the pattern of two-dimensional (2D) filters are manifest to check; however, one-dimensional (1D) filters are relatively abstract to be inspected in a straightforward manner. For example, if 1D filters are plotted out, only peaks and troughs can be directly spotted. Therefore, it is difficult to have intuitions upon investigations in the time or spatial domain. Alternatively, by investigating the power spectra of the learned filters in the frequency domain, certain conclusions could be drawn by studying their impacts or modulation on EEG data.

The frequency Ω of unit radian per second for the EEG data in continuous case is associated with ω of unit radian per sample in a discrete case by $\omega = \Omega T_s$, where T_s is the sampling frequency of 250Hz. Because the EEG data underwent a bandpass filter with the highest frequency equal to 50Hz, the highest angular frequency Ω_h is $2\pi * 50$; and $\omega_h = \Omega_h T_s = 2\pi * 50/250 = 0.4\pi$. Hence, when utilizing DFT to convert the learned filters from the time domain to the frequency domain via formula $H(e^{j\omega}) = \sum_{-\infty}^{+\infty} h(n)e^{-j\omega n}$, the range that needs to be focused on is $[0, 0.4\pi]$. Because in theory there should be hardly any signal beyond 50Hz, or severely attenuated one if exists.

Considering the bottom layers are inclined to learn concrete features, the learned filters of the second hidden layer are taken for analysis. The 512 filters of this layer incur the impracticability to inspect them individually. Therefore, the filters' collective behaviours are investigated by applying clustering to them. In detail, the filters are clustered into eight categories using Euclidean metric, with the corresponding number of elements for each category showing in Fig. 3.11 as legend. Simultaneously drawn in Fig. 3.11 are the frequency amplitudes of the cluster centroids of each category. Both angular frequencies in different units are labelled on the x-axis too.

As in Fig. 3.11, the solid line represents the cluster with the most elements, which spans across the upper alpha band and almost beta band. The dashed line represents the cluster with the second most elements, which spans across the alpha band and slightly partial beta band. These distributions coincide with the laboratory's

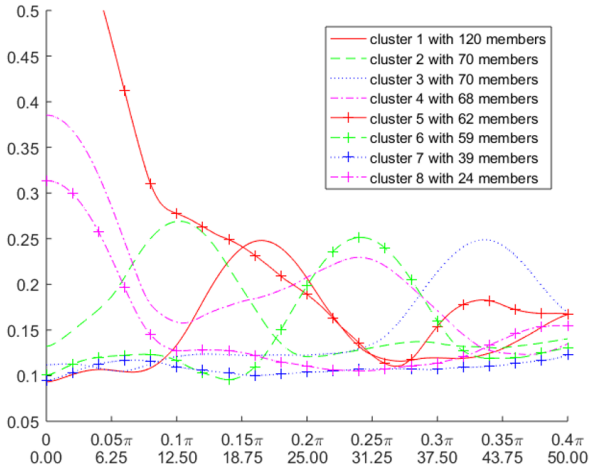


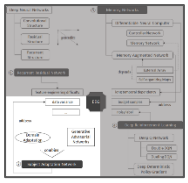
Fig 3.11 Clustering of frequency amplitudes of learned filters.

previous research [70, 94], peers' research [95] and general neuroscience recognitions [96]. For the beta band, the postulate tends to emphasize its correlation with mind state such as vigilance [97]. Alpha band is usually linked to visuomotor condition [98, 99]. According to the design of the experiment, these two bands should play critical roles in unveiling the test subjects' mental states and behaviours. Consequently, it is enlightening to know that what was automatically learned agrees with what to be expected.

The category with the least members is represented by a dash-dot-plus line. Roughly speaking, it acts as a low-pass filter. However, it is argued that lower frequencies, such as delta band, are not that useful in most scenario and some research just put it out of consideration [71]. It is also interesting to see that the auto-learned filters with least numbers favour such a fact. In addition, there is difficulty in explaining the dot-plus line which represents the category with second least members. Overall, it is encouraging to see that different filters focus on different sub-bands and together they cover all the available bands.

It is also noticed that during hyper-parameters tuning process, the filter length is not very critical to the performance of the model. An explanation from the signal processing perspective is, the properties of the filters, e.g., low-pass filters, bandpass filters, etc., are intrinsic to the targeted problem itself. The neural network is trained to have these filters emerged. Alternatively speaking, if the optimal filters should be low-pass filters, the neural network is deemed to learn these low-pass filters instead of high-pass filters. It is reminded that to design a filter in the time domain, the corresponding version in the frequency domain with specified property will be examined first. The length of filter in time domain, i.e., the number of coefficients, only decide the detail or fineness of the filter, but not the fundamental property. This indicates that it can be more leeway in choosing the filter length, especially for very deep neural network. At last, it might be worthwhile to mention that, to design filters jointly covering the whole frequency spectrum of EEG data, may potentially favour better analyzing results, as indicated by Fig. 3.11.

4. Variance Reduction by Adversarial Method



⌖ This chapter covers the bottom-left region of the research map in Fig. 2.8. By combining the concepts of domain adaptation (DA) and ideas of generative adversarial networks (GAN), this chapter designs and implements a special DNN resembling the architecture of GAN in the context of DA, to reduce the intra-subject and cross-subject variance of EEG data .

4.1 Background

As mentioned in section 2.2, although EEG is a convenient and economic way to inspect brain activities, it tends to display distinct intra-subject and cross-subject variance due to the individual physiological traits. Cross-subject variance is intuitive to understand. Hence, intra-subject variance is discussed here by taking the driving experiment in section 2.3.1. In Fig. 4.1, it shows the RTs of one subject that participated in the experiment for two times. According to the experiment log, these two experiments were carried out with identical setup, similar time of the day, and similar physical and mental conditions subject reported when participating. Nevertheless, the discrepancy between RT patterns for different sessions of the same subject is eminent, which is an indirect indicator of intra-subject various of EEG data.

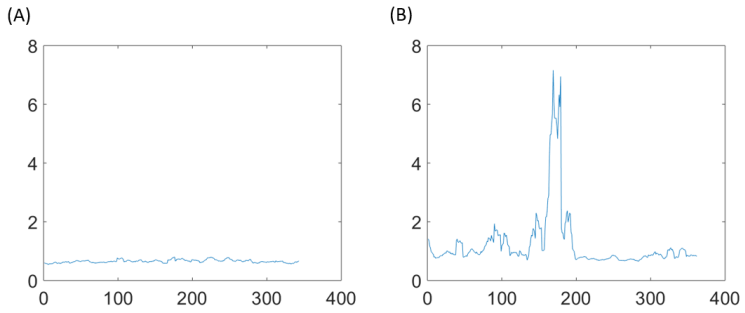


Fig. 4.1 RT patterns for the same subject who participated in the experiment for two times. The x-axis indicates the sequence of trials during the experiment; The y-axis indicates lengh of RT (unit: seconds).

Such discrepancies pose several challenges for the performance of conventional methods. Considering the case that the drowsiness of the subject is of interest, however, EEG signals from the same subject in different runs of the experiment or different subjects with similar states tend to present different power spectra. Reciprocally, identical EEG patterns can be labelled quite differently especially for different test subjects. Such inconsistencies have broad negative impacts on EEG signal processing, from quantitative analysis to qualitative identifications. This problem is even worse for applications because new subjects usually have new distributions different from their counterparts used for prototype method build-up.

This problem is not peculiar to EEG itself indeed. It has been observed and addressed in conventional machine learning studies, and the corresponding techniques are termed transfer learning (TL) or domain adaptation (DA) [100-104]. The necessity for such a long-standing research is recapped in Fig. 4.2. Suppose a model is constructed to target a binary classification problem. In Fig. 4.2(A), a training process leads to a decision boundary which is nicely set up between the two training class

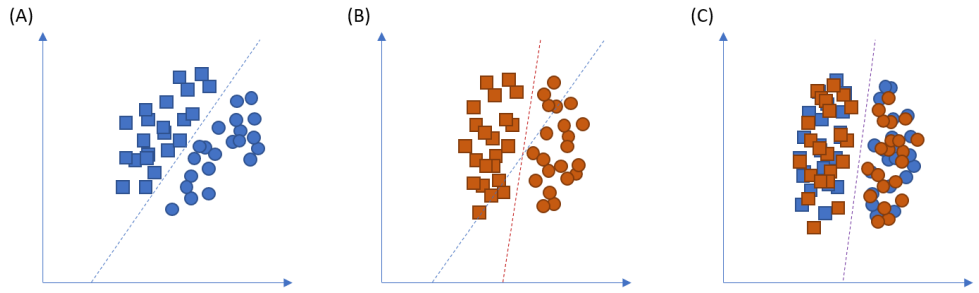


Fig. 4.2 (A) Distribution of training samples and the corresponding decision boundary; (B) Distribution of test samples and the corresponding decision boundary as well as migrated decision boundary from training; (C) Adapted training samples and test samples and corresponding decision boundary. Note the coordinate in (C) is not necessarily the same as (A) and (B), because the adapted samples can be in some latent space.

samples. However, due to the discrepancy of training sample distribution and testing sample distribution, simply migrating the decision boundary obtained in training might be inappropriate for the test set, as illustrated in Fig. 4.2(B). For binary classification, it is assumed that both the training set and the test set comply with some bimodal distributions. Hence, by embedding both into some latent space, where the same class of training samples and test samples are with more coherent distributions, the generalisation of the trained model is certainly with higher performance, as illustrated in Fig. 4.2(C).

However, the difficulties facing EEG lie in the following several folds: (1) EEG data are immensely complicated and generally of high dimension. Not to mention the difficulty to get certain intuitive estimations such as distributions in different domains, e.g., the training set domain and the testing set domain; even to glean apparent information from different subjects in the same domain is still a challenge. (2) DA usually works well for well-chosen features obtained from original data, but it is generally not applicable for EEG signals due to the limited understanding of the underlying neurophysiological/cognitive processes. It tends to be disputative that what kind of features extracted from EEG data suitable for a particular neurological experiment. (3) For some scenario such as classification, the ground-truth labels for the corresponding EEG data are sometimes also problematic. For example, in the driving fatigue experiment in section 2.3.1, the drowsiness is usually indirectly measured in an indirect manner, such as by RT. However, this method unavoidably introduces noise to the labels, which incurs the ineffectiveness of applying DA method via labels.

Work in [44] is regarded as a precursor for adaptation from DL perspective. It tries to solve the problem with the guidance of GAN and harnessing DL's end-to-end modelling philosophy. The work harmonically combines several components together to achieve adaptation. Essentially, the basic idea in [44] is to seek some common representation form for both domains. Later, methods such as associative domain adaptation [105] introduce more appropriate loss measurement and generalize the previous work in [44]. Another approach to utilize the neural network for DA is making use of GAN to transform one domain into the other, e.g., the source domain to the target domain and/or vice versa, as in [106, 107]. Since the objective is now clearer than the case of finding a common representation space that is not straightforward, ADDA and CycleGAN in [106] and [107], respectively, have achieved more astonishing results.

In addition, the achievements in [44] make it appealing to use the adaptation techniques built upon neural networks for EEG signal processing. However, one prominent challenge for EEG data is besides the intra-subject variance, the cross-subject variance has a smaller granularity than the original DA problem. In later section,

a subject adaptation network (SAN) is introduced to specifically target EEG signal processing as rooted. It is pointed out that in a narrow sense, it can also adapt EEG data from different experimental sessions of the same subject; nevertheless, the term subject adaptation network is used instead of session adaptation network to denote its generality.

4.2 Subject Adaptation Network

4.2.1 Mathematical Description

The EEG signal is one of the most complex bio-signal in neurophysiological or cognitive research. Comparing the EEG signal processing with conventional problems such as image classification, two criteria have to be met for a neural network model to perform well: (1) it must be capable of automatically extracting features to solve the problem such as classification in the first hand. (2) the extracted features can be shared between the subjects. The first criterion is justified by the achievements of deep learning. For the second criterion, various methods have been proposed and devised recently as aforementioned. However, it is still worthwhile to guide the design of the SAN via deductions from the theoretical perspective.

For a given cognitive experiment especially with BCI oriented, the set of subjects who participated in the experiment is denoted by $\{s_i\}_{i=1}^{N}$. Usually a subject s_i participated in the experiment for several times (or sessions), with each session comprising of many experimental trials (or epochs). For analysis, the epoch data for a specific subject, s_i, are aggregated and denoted by $\left\{s_i^j\right\}_{j=1}^{T}$. T denotes the total number of trials over all sessions involving subject s_i. There is also a corresponding label for each trial in general. Taking the P300 experiment in section 2.3.1 as an example, each trial s_i^j corresponds to a target or a nontarget stimulus. For convenience, the pair (x, y) with $x \in \left\{s_i^j\right\}_{j=1}^{T}$ and y (the corresponding label) is termed input/label. In the following if there is no ambiguity, x is implicitly assumed to be always coupled with the label y. Suppose the probability density function (PDF) of $\mathbb{P}_{s_i}(x, y)$ is $p_{s_i}(x, y)$ (denoted as $p_{s_i}(x)$ in the following for brevity) in the original data space (or sample space); then the cross-subject variance means there is an obvious discrepancy between $p_{s_i}(x)$ and $p_{s_j}(x)$ for the different subjects s_i and s_j with the same label y.

To reduce the variance, one obvious idea is to find another space (called feature or embedding space) in which the transformed PDFs better align with each other. Defining the mapping from the data space to the feature space by L, an optimisation problem can be formulated as follows.

For $x \in \left\{s_i^j\right\}$, let $z = L(x)$. Suppose $x \sim p_{s_i}(x)$, the corresponding distribution of z is denoted by $q_{s_i}(z)$, aka $z \sim q_{s_i}(z)$. Based on the denotation above, criterion (2) can be formulated as optimisation of the following formula:

$$\underset{L}{\operatorname{argmin}} \int \max_{i}\left\{q_{s_i}(z)\right\} dz \tag{4.1}$$

$$q_{s_i}(z) = p_{s_i}\left(L^{-1}(z)\right)|1/L'| \tag{4.2}$$

An intuitive illustration of (4.1) is shown in Fig. 4.3. Assume that there are two subjects in the dataset denoted by s and t respectively. s is for training and t is for testing. After transforming from the data space to the feature space, the corresponding distributions are denoted by $q_s(z)$ and $q_t(z)$. Suppose some model is trained using s

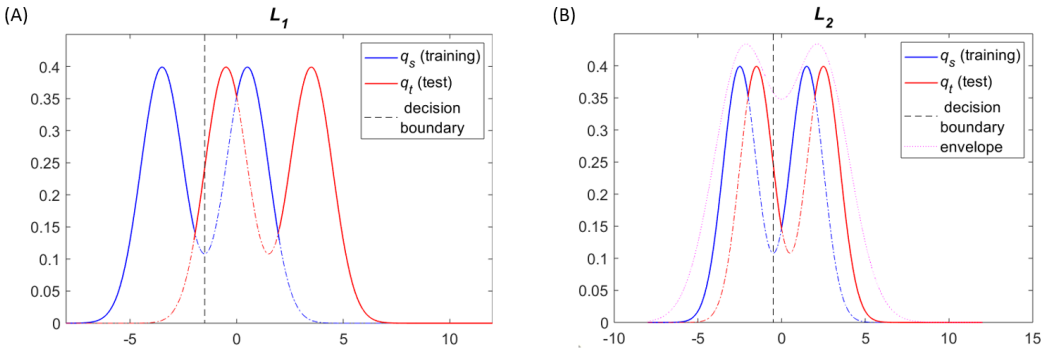


Fig. 4.3 Illustration of the transformed probability density function alignment. L_2 is preferable to L_1 because the area under the solid line of (B) is smaller than that of (A), which merits the criterion formula (4.1).

and made the inference on t. For different transformations L_1 and L_2, suppose the learned decision boundaries are those shown in Fig. 4.3. It is obvious that the generalization in Fig. 4.3 (B) is better than that in Fig. 4.3 (A). The more coherent alignment of the transformed density functions in Fig. 4.3(B) result in a lower value of the integral $\int \max\{q_s(z), q_t(z)\}\, dz$; consequently, there is a greater preference for the corresponding transformation L_2, which corresponds to better generalization.

However, the optimization of (4.1) is a variational problem from a mathematical point of view. Because the potential feature space and the form of L are unknown, the problem is generally highly intractable. Nevertheless, observation of the transformed PDFs and their alignments can suggest some characteristics of the optimal L which in turn guide the design of the neural network. For example, it is obvious that an envelope can be found for the aligned PDFs as indicated by the dotted magenta line in Fig. 4.3(B). If the original distributions of subjects are bimodal, it is expected that the envelope has similar statistical property. Or equivalently, L should be endowed with some modality preservation property.

4.2.2 Network Architecture

As mentioned above, the data distribution $p_{s_i}(x, y)$ is usually inconsistent with $p_{s_j}(x, y)$, for subjects participating in the same experiment with an identical setup, or even the same subject participating in the same experiment in multiple runs (now s_i and s_j are referred to the data from different sessions of the same subject). Such inconsistency in samples is the root cause of difficulty in EEG data analysis. A desired solution is to enforce adaptation to all subjects' data. However, as unveiled above, theoretical formulation into an optimization problem does not mean that an analytical solution can be found in practice. For example, a 32-channel EEG data with sampling rate 250Hz is 8000-dimensional, which prohibits any direct observation. The interpretation of the original distribution is undoubtedly challenging, never mention the evaluation of the adapted distribution.

However, distributions in low-dimensional spaces such as 1D and 2D are much easier to interpret and manipulate. In the previous section some indications have been drawn about the transformation L and the post-adaptation distributions. Based on certain domain knowledge of EEG experiment, it is appealing to design an artificial distribution in a low dimension and enforce the original sample distribution to approximate such an ideal distribution endowed with nice properties. Thus, the intuition and motivation for this work are to avoid gleaning the distributions of samples in a very

high-dimensional space, and just enforce alignment with a "clean distribution" in low-dimensional space of the same modalities by an adversarial network.

4.2.2.1 Desertion of Network Components

The reasoning and insight above lead to the network architecture designed in Fig. 4.4. The overall architecture consists of a generative and a discriminative network, resembling the general architecture of GAN [42]. The generative network or generator is split into two parts, namely, an adapter network denoted by $A(x, \theta_a)$ and a mapper network denoted by $M(x, \theta_m)$. The discriminative network or discriminator is denoted by $D(x, \theta_d)$. After adaptation, the network can aggregate other components such as a classifier network $C(z, \theta_c)$ or a sample selection module $S(z, \theta_s)$ for post-adaptation stage applications, as explored in the following experiment section.

However, this work is different from the original GAN in two aspects. First, the input to the generator here is based on the sampling of real data, instead of from a random source as in the case of GAN. Hence, rather than learning a function that maps the input distribution to a target distribution, the generator learns to align distributions from different sources into a coherent one to confound the discriminator. Second, the discriminator receives the ground-truth information not from samples in the real world, but by sampling a designed or artificial distribution.

The data flow of the proposed model is as follows: EEG signals either from different sessions or from different subjects are input into the generator for distribution alignment. The dimensionality-reduced latent representations from the generator are pipelined to the discriminator in competition with another discriminator input, which is by sampling an artificial distribution. The adaptation of EEG signals is guaranteed by the working principles of GAN during the training process, and these adapted representations are harnessed for different applications, as illustrated in Fig. 4.4.

It is noticed that the generator here is split into an adaptor plus a mapper. One reason is that when the adaptor tries to project the original sample data into some embedding space, it still can keep the projection in reasonable dimensionality during the adversarial learning process. Such a dimensionality is mandatory to have later applications such as classification to perform well. Another reason for splitting the generator lies in the fact that the "real-" distribution is artificial. Designing such a distribution requires domain knowledge of and insight into the original sample space. However, if there exists some biased design for the target distribution, the enforced

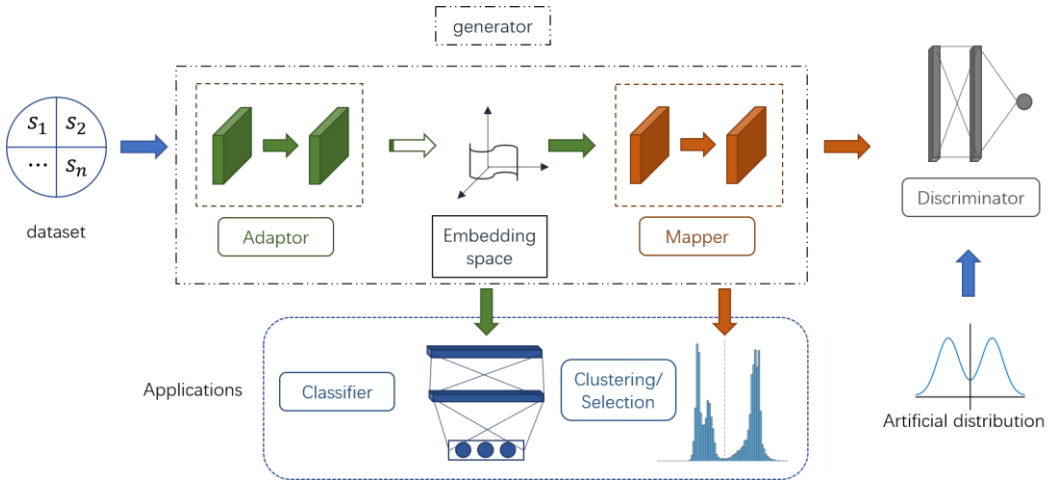


Fig. 4.4 The overall architecture of the subject adaptation network.

final distribution, aka output from the generator, might be problematic. By splitting the generator into an adaptor and a mapper, it is intended to harvest only intermediate transformed representations that have the inclination to be better aligned. Furthermore, as mentioned, directly processing the final output of the generator might be inappropriate for some applications such as classification via a deep neural network. The intermediate representations still have reasonable dimensions, which can aid the classification performance. Nevertheless, determining the boundary between the adaptor and the mapper is still an empirical process.

Based on the above reasoning, before processing the subject data, samples from all subjects will go through the adaptation network which is designed to align the distribution of subjects while still maintaining internal distinguishability. It is expected that by competing with the artificial targeted distribution, data from different subjects can be coherently and consistently aligned while still keep the modality for later processing such as classification. Consequently, it is clear that the paradigm of utilizing the proposed model for EEG data analysis consists of two stages. The first stage is to train the adaptor, mapper and discriminator to the optimal balance, which means that the discriminator cannot effectively determine the sources of its inputs. The second stage is to pipeline the adaptor with or without the mapper to other components according to different applications.

4.2.2.2 Artificial Target Distribution Design

Generally, it is impossible to directly observe the distribution of EEG data in the sample space due to the tremendously high dimensionality of the data. However, some overall properties of the distribution such as the modality and relative size of the potential clustering of the original data can be depicted. For example, the P300 EEG experiment in section 2.3.2, subjects react to two kinds of stimuli, targets and nontargets. It can be expected that there are two modalities for the designed artificial PDF. If the ratio between targets and non-targets is supposed to be 1:2, it can be further assumed that the area under one modality is half of the area under the other modality, as indicated in Fig. 4.5.

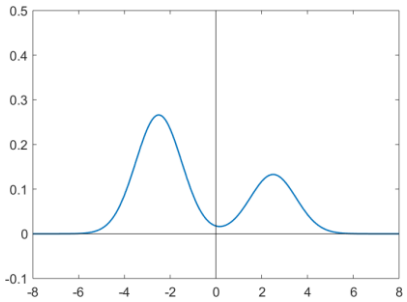


Fig. 4.5 Designed artificial distribution for an exemplified EEG dataset.

To feed the discriminator by sampling from the artificial distribution, in this work the rejection sampling is always used considering its simplicity and efficiency in low dimensions.

4.3 Experiments and Analysis

4.3.1 Driving Task EEG Dataset

As aforementioned, EEG can reflect certain characteristics of mind state or cognitive process pertaining to people's specific activities such as driving. By

collectively analyzing samples from the test subjects, some conclusions especially from the neurophysiological perspective can be reached, such as activities of separate cortices, connectivity between different lobes, etc. However, the difficulty in EEG research is how to select the most appropriate samples for analysis, to correlate such implicit mind states with the brain activities. For example, to understand the brain dynamic under low performance driving, one needs to screen out the trial data corresponding to fatigue for analysis.

In this experiment, EEG dataset captured during the experiment in section 2.3.1 is utilized to demonstrate the capability of the proposed network for sample selection, especially from the intra-subject variance perspective. In the simulated driving experiment, RT is utilized for labelling fatigue level. But usually the relation between the direct but implicit label such as fatigue level, with indirect but explicit indicator such as RT, is unknown. And due to other factors, the measurement is usually error-prone or at least accompanied by noise. It can potentially move the EEG data indeed corresponding to the alert stage into the drowsy category; consequently, it may hurt the objectiveness of the conclusions. Actually, even for one subject, the higher number of runs the subject participate in the experiment, the more likeliness of inconsistency on data labelling. Such intra-subject variance potentially poses a challenge for effective analysis.

This experiment targets the intra-subject variance challenge for better sample selection. So, one subject who participated in the experiment several times with the greatest number of trials is chosen for demonstration of the proposed model utilization. The EEG signals are down-sampled from the original 500Hz to 250Hz, then a band-pass filter with range 0.5 to 50Hz is applied. After converting from the time domain to the frequency domain via FFT, just the alpha band (4~7Hz) of the frequency spectrum is selected to demonstrate the key steps. The alpha band data are converted to topographical images as in Fig. 2.1.

For a specific sample, the measured RT is potentially biased. For example, due to subjects' distractions or weariness of muscles, the prolonged measured RTs probably brings a sample corresponding to alertness into the category of fatigue. However, it is believed that the overall trend of RT is trustworthy. Therefore, to design the artificial distribution, the first step is to plot the histogram of measured RTs to get an overall estimation of the distribution. Fig. 4.6(A) is the histogram of RT; hence, a gamma distribution is chosen in Fig. 4.6(B) as the designed distribution.

With the specified configuration in Table 4.1, the EEG data are directly processed in time domain by training the network for 40000 iterations with a learning rate 0.0005 until the enforced alignment aka the gamma distribution is clearly observed. To maintain stability during training, the model takes advantage of the loss suggested

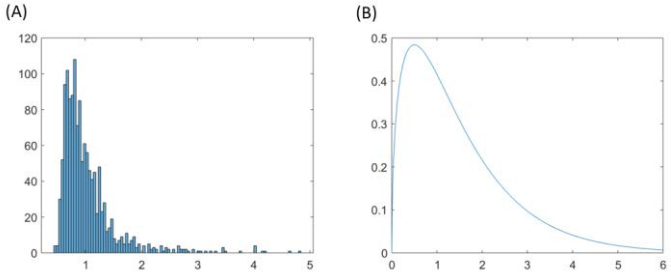


Fig. 4.6. (A) Histogram of measured RTs; (B) Designed gamma distribution approximating the distribution of measured RT (alpha=1.5, beta=1.0).

Table 4.1 Network configuration for driving dataset

Name		Layer	#Filter	Kernel	Activation
Adaptor	L1	Conv2D	32	3x3	-
	L2	Conv2D	32	3x3	-
		AvgPool	-	2x2	-
	L3	Conv2D	64	3x3	-
	L4	Conv2D	64	3x3	-
		AvgPool	-	2x2	-
Mapper	L1	Linear	256	-	sigmoid
	L2	Linear	1	-	-
Discriminator	L1	Linear	256	-	ELU
	L2	Linear	64	-	ELU
	L3	Linear	1	-	-

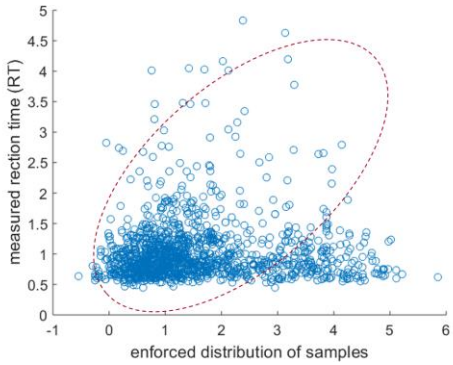


Fig. 4.7 Plotting of enforced sample distribution vs. measured RT. The dashed eclipse indicates the recommended range for selecting samples.

by Wasserstein GAN (WGAN) [108] as well as spectrum normalisation [109]. The enforced distribution of samples vs corresponding RTs is plotted in Fig. 4.7 to select the most appropriate samples for further analysis. In Fig. 4.7, according to our domain knowledge, the orange dashed eclipse is plotted to indicate the recommendation of EEG samples for corresponding analysis, because the measured RTs are more consistent with the enforced aligned distribution diagonally.

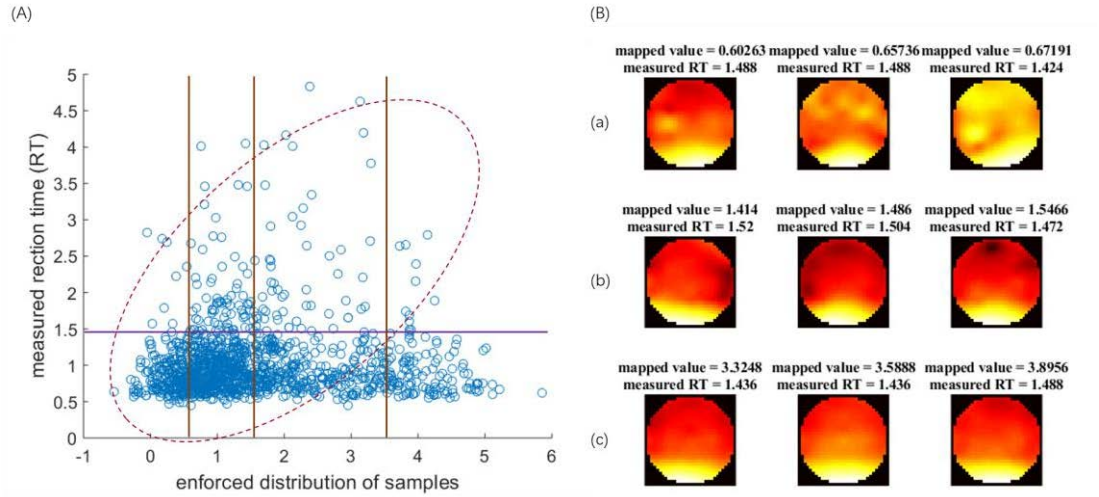


Fig. 4.8 Topographies of the samples in different groups. (A), (B) and (C) are chosen based on the measured RT and mapped values of the enforced distribution.

To demonstrate such a claim, the topographies of neighbouring samples in three groups are inspected along the line $y = 1.5$ in Fig. 4.8. It can be noticed that the topographies from the same group are sharing more similarities, which can distinguish them from other groups. Hence it justifies the claims as well as the suitability of utilizing the model for sample selections.

The procedures for performing sample selection based on SAN are summarised as follows:

Estimate the measured or noise label distribution.

Design the artificial probability density function f for measured labels.

Train the network to obtain the optimized mapping function $L(x; \theta^*)$ (generator network).

Scatter-plot the distribution of network outputs $L(\{x_i\}; \theta^*)$ vs. measured labels.

Propose the range R based on domain knowledge.

Only select samples in range R for analysis.

4.3.2 Oddball Task EEG Dataset

To demonstrate the performance improvement obtained by adapting EEG data of different subjects for further analysis such as classification, an EEG dataset captured during the VEP oddball task in section 2.3.2 is utilized here. EEG data is usually processed in the frequency domain; however, considering the feature extraction capabilities of DNN, this experiment directly works with waveform EEG data in the time domain.

For data preparation, data from four subjects are first rejected due to corruption or poor responses. Next, the EEG signals are band-passed to 1-50 Hz, followed by down-sampling to 64 Hz. Then epochs are extracted within [0, 0.7] second interval time-locked to the stimulus onset. Epochs with incorrect button press responses are excluded to reduce the impact of outliers. Finally, the mean baseline is subtracted from each channel in each epoch for normalization. This results in 382 targets vs 1146 nontargets. To counter the potential bias during training, the imbalance dataset is resampled to have a ratio of targets to nontargets as 1:3, which is also taken into consideration when the artificial distribution is designed, as in (4.3):

$$y = (0.5 \mp 0.25)/\sqrt{2\pi} * \exp(-(x \pm 2.0)^2/2) \tag{4.3}$$

Table 4.2 Network configuration for Oddball dataset

Name		Layer	#Filter	Kernel	Activation
Adaptor	L1	Conv1D	32	7	-
		Downsample	-	2	-
	L2	Conv1D	64	5	-
		Downsample	-	2	-
	L3	Conv1D	128	3	-
		Downsample	-	2	-
Mapper	L1	Linear	64	-	tanh
	L2	Linear	1	-	-
Discriminator	L1	Linear	256	-	ELU
	L2	Linear	64	-	ELU
	L3	Linear	1	-	-
Classifier	L1	Linear	256	-	ELU
	L2	Linear	2	-	-

Fig. 4.4 suggests that during adaptation, the outputs of intermediate layers can be in more coherent alignment than the data in the original sample space, while they still retain reasonable dimensionality for classification. The relation between the embedding space and the classifier is highlighted in Fig. 4.4. Noticing that the dimension of the sample space is 2880 (64 channels with each channel having 45 data points). These limited samples in such a high-dimensional space are far from enough to capture the underlying data distribution. The available data constraint incurs the choice to practically analyze only one single channel which is Pz, to restrict the sample space to 45 dimensions. Based on this limitation, a network configuration is given in Table 4.2.

With the targeted distribution, the network is trained with the Adam optimiser of learning rate 0.0001 for 10000 iterations. The enforced distribution with ground-truth labels is shown in Fig. 4.9. Due to the complexity of EEG data, the separation is not as clear as the case of MNIST data.

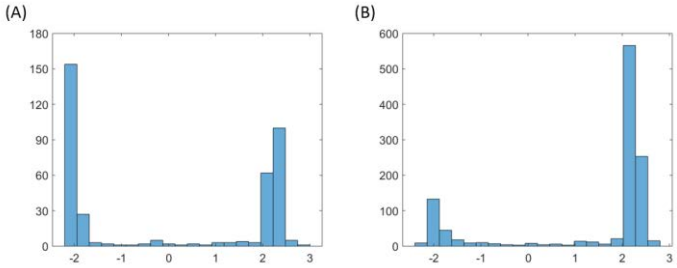


Fig. 4.9. (A) Histogram of samples corresponding to target (B) Histogram of samples corresponding to non-target.

For benchmark comparison, the conventional support vector machine (SVM) method and the current well-spoken-of EEGNet are chosen [56]. Because the data is just of single-channel, the depthwise convolution which is analogous to the common spatial pattern (CSP) filters [110] is omitted from the original EEGNet structure. For the sake of brevity, only the first five subjects are utilized for leave-out testing.

The results are listed in Table 4.3. With budget EEG data, all models produce comparable results, but the proposed model is slightly better among all. It is mentioned that SVM has a big advantage with limited training samples, because just a few support vectors are enough to determine a decision boundary with reasonable margin. But it is believed that by increasing the number of samples to a reasonable magnitude which is appropriate to train the proposed network in the adversarial manner, the improvement could be further boosted. Nevertheless, the improvement via subject adaptation can still be justified from the results, even with limited samples.

Table 4.3 Test results comparison

Model	S1	S2	S3	S4	S4	Average
SVM	0.815	0.805	0.789	0.811	0.780	**0.800**
EEGNet	0.808	0.900	0.806	0.778	0.732	**0.805**
SAN	0.802	0.885	0.820	0.809	0.760	**0.815**

The procedures for performing classification based on the SAN are summarized as follows:

Estimate the sample distribution.

Design the artificial probability density function f that simulates the estimated sample distribution.

Train the network to obtain the optimized mapping function $L(x;\theta^*)$ (generator network).

Empirically choose some intermediate latent representations of the generator (which are equivalent to the outputs of the adaptor) $A(x;\omega^*)$, here $\omega^* \subset \theta^*$.

Train another classifier $C(z,\vartheta)$ based on the outputs of the adaptor.

4.4 Discussion and Future Work

Some tricks recognized as subtle and useful are highlighted here to help ease the subsequent research.

First, by observation, average pooling is preferable to maximum pooling when building the network especially for the adaptor. Due to the univariance effect, max-pooling may affect the distinguishability between samples because the details are blurred to some extent. One consequence is the potentially induced higher variance as shown in Fig. 4.10. Another consequence is the potential failure of adaptation when the variance exceeds a certain width, which means the alignment cannot be effectively enforced.

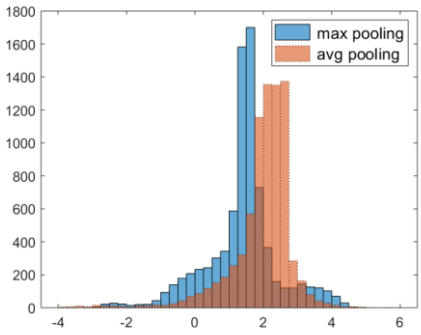


Fig. 4.10 The impact of different pooling methods on the enforced distribution of samples.

It is manifest that the properties of designed distribution should be taken into consideration for the choice of the mapper's activation function. For example, in the first experiment, the sigmoid function is chosen since gamma distribution only exists when $x > 0$. For the second experiment, the hyperbolic tangent function is chosen because the designed target distribution is to some extent symmetric with respect to $x = 0$. It is also noticed that the mapper is quite sensitive to initialization. When Gaussian normalization is used, the standard deviation is suggested to keep within 0.5. Using larger values tend to cause modal collapse. One possible reason might be the initial penalty is severe that it restricts the exploration throughout the whole training process. The comparison is demonstrated in Fig. 4.11.

It is also found that it is slightly easier to use the WGAN framework instead of the original GAN for adversarial training, because it allows an initially relaxed exploration of the parameter space. Nevertheless, once the nearly optimal parameters

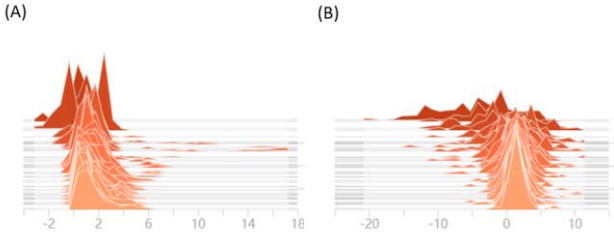


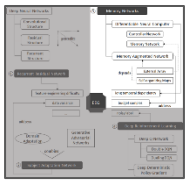
Fig. 4.11 Enforced distribution trends during training with different standard deviation for the second experiment (A) $\sigma = 0.2$ (B) $\sigma = 1.0$.

were found and marginal fine-tuning commences, the selection between these two frameworks does not make much difference.

One restriction for EEG data analysis is that EEG datasets are usually costly to obtain. However, the use of the proposed SAN method requires to sample the original data space in a reasonable amount. However, the availability of EEG samples can satisfy only part of the budget requirement for adversarial training. It is expected in future work by considering the proposed model in a wider sense, harvesting its potential to generate additional interesting outcomes could be fulfilled.

The fact that the performance of the proposed SAN relies on a properly designed artificial distribution whose modalities and relative shapes and distances between modalities can well reflect the original sample distribution, provides another aspect for future work. One idea is to utilize principal component analysis (PCA) on a subset of samples, to map these samples into a lower dimensional space. Then using GAN to learn a generating network which simulates the distribution of samples by competing over samples from the latent sample space after PCA. However, studies of these topics are just launched, and more efforts are needed for deeper investigations in these areas as future work.

5. Memory Network Perspective



⌖ This chapter covers the top-right region of the research map in Fig. 2.8. By studying the work of differentiable neural computer (DNC), this chapter proposes the stacked DNC and to use self-organizing maps (SOM) as memory module, to increase the capability of unveiling temporal correlation of EEG data and to mitigate the budget EEG data constraint.

5.1 Background

Research of neuroscience indicates that two components in human brain play key roles in human intelligence, i.e., the memory elements and the interconnections between them [111-113]. Although with huge successes in a variety of applications, the current prevailing DNN structures still fails this neurobiological perspective and are criticized as mixture of memory and computation.

Taking image classification for example, the weights of convolutional layers are generally regarded as the representation of the computation process. Moreover, convolution might be treated as template matching; in this way, the weights can also be interpreted as the knowledge of objects to be classified. However, these interweaving interpretations of computation and knowledge post challenge in explaining what have learned. Another drawback is that the blur boundary between computation and knowledge incurs the difficulty in increasing the capability of the network. For example, increasing memory will unavoidably increase the computation cost. There is no way for a neatly isolation of one perspective from another in general.

There were attempts at investigating the interaction and integration between memory and computation along the history of neural networks [114]. Recently, Google proposed a memory network called differentiable neural computer (DNC) in systematically exploring external memory to aid the computation of conventional neural networks [78, 115-117]. The ingenuity was the establishment of an enriched computational model by analogous to modern computers. Although it shows amazing performance on various difficult tasks, the utilizing memory of DNC avoidably adds the complexity to the whole architecture and makes the adjustment or modification to the original system extremely difficult.

Considering that Google and Facebook already begun the work with promising achievements, it is worthwhile to carry out further investigation from various aspects, such as the way that brain works, to seek further accomplishments. Moreover, memory network is an important computational intelligence model, hence utilizing and extending memory network for EEG data analysis is a meaningful work.

5.2 Stacked Differential Neural Computer

The motivation for the invention of DNC is natural from the neurobiological perspective. It is exemplified by the fact that the high cognitive level of human beings is greatly attributed to the brain's memory mechanism, based on which tasks from simple recall to complex reasoning could be well performed. It is also believed that this new memory-augmented architecture which has already demonstrated its remarkable capability in solving complex tasks would lead to new horizons [117].

In this section, the operation and memory mechanism of DNC are investigated and extended. Based on these studies, a stacked DNC architecture is proposed for EEG data analysis. In fulfilling this goal, it is necessary to have a good understanding of DNC first to exert its potential from an application perspective. When tracing the evolution of DNC, from its ancestor, i.e., the neural Turing machine (NTM) [115], to its current state, it can be well understood that the interaction with memory is the most critical and complicated part of the whole system. It is better to begin with the approach of memory interaction in order to have an essential interpretation of DNC. Hence, an intuitive interpretation about the way of memory augmentation in DNC is necessary and intriguing for subsequent work.

5.2.1 Interpretation

Recurrent neural network (RNN) especially with the LSTM cells is currently considered the most successful model in utilizing memory mechanism [89, 118, 119]. Unlike plain multi-layer perceptron (MLP) network, the recurrent structure endows it with the ability to recall or correlate the past input with the current one. The RNN and its unfolding paradigm across the time axis are illustrated in Fig. 5.1(A). Generally, the hidden state at the previous time step $t-1$ is concatenated to the input at current time step t for further processing during the recurrent procedures. Usually, it is presumed that the temporal correlation exists between successive inputs. It is also the motivation to apply RNN. However, it might not always be the case, such as in the copy task [78]. The authors believe that what should be emphasized is not the temporal pattern in data, but a way of programming.

However, due to the characteristic of the RNN structure emphasized in Fig. 5.1(B), the direct link between the adjacent time steps unavoidably incur temporal couplings. The advantage is that it can capture the correlation if it does exist for the

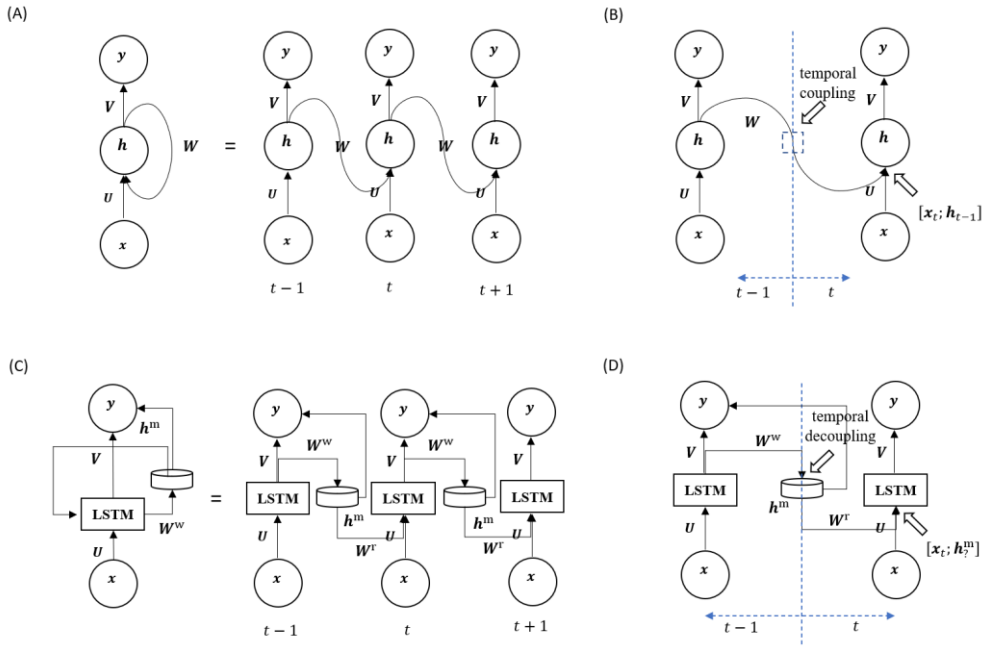


Fig. 5.1 (A) RNN and its equivalent unfolded version along the time axis (B) Demonstration of the tight temporal coupling in RNN (C) Equivalence of the unfolded DNC across the time axis (D) Demonstration of the temporal decoupling by augmenting external memory. Note x, y and h are for input, output and hidden state respectively. Capital letter U, V and W are used to denote weight matrices. For denotation $\boldsymbol{h}_{?}^{\mathrm{m}}$ in (D), the question mark indicates that due to asynchronous operation, the hidden state at which step (latent time step) involved in the current operation is unknown in general.

problem being modelled. The drawback is that such a tight coupling can be futile in certain situations. For instance, if the coupling to the current state is some hidden state that several time steps prior to the current state, it is probably faded away when passing through these links up to the current time step.

By the introduction of LSTM, such a restriction is relieved to some extent. However, even a shallow reflection of human's cognitive processes can reveal the fact that information or experiences obtained a long time ago, are probably involved in the problem currently being pondered. Hence, a natural question to raise is that whether such a temporal coupling (weak in the case of LSTM) in traditional RNN can be mitigated or even eliminated, but also re-acquired when necessary.

As a comparison, the DNC and its equivalent unfolded structure are illustrated in Fig. 5.1(C). Compared with RNN, a matrix or cell block exists acting as an external memory (indicated by the disks in the figure) which characterizes the speciality of the architecture. Intuitively speaking, the common link between successive unfolded hidden layers is squeezed by memory sitting in-between. There exist mechanisms of reading from and writing to the memory at each time step t. This modification incurs the complexity from spatial perspective; however, the temporal dependency is dramatically reduced due to the memory access methods. The links and interactions between unfolded hidden layers of DNC's controller are shown in Fig. 5.1(D), by means of which coupling can be eliminated and regained when necessary, especially when DNC is tactically trained.

Based on the interpretation and comparison above, an intuitive interpretation for DNC could be stated as that it is targeted to solve the temporal coupling in conventional RNN. Such an interpretation is preferable than some exceedingly general ones. One hand is that such an insight indicates any method which can modulate the temporal coupling in RNN can be put into consideration. On the other hand, from application perspective, problems with an implicit dependence between long distance inputs, or weekly dependence between consecutive inputs can be considered to utilize DNC for modelling.

5.2.2 Network Architecture

The above interpretation leads to a natural construction to stack multiple DNCs together, to form the stacked DNCs. This arrangement just resembles the multi-layer RNN or LSTM network which are ubiquitous in DL applications. A direct comparison is given in Fig. 5.2. The reason that in Fig. 5.2 for stacked DNCs, the recurrent path and external memory are both in dotted lines, is to emphasize that they are internal to DNC instead of external. The current drawing is for analogous purpose.

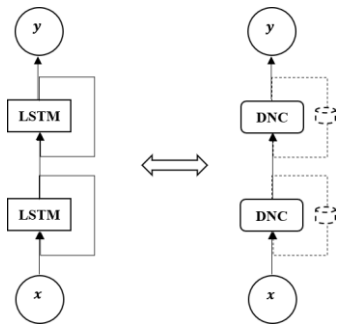


Fig. 5.2. Comparison of multi-layered LSTM and DNC.

There are several advantages to stack DNCs together. The most obvious one is by doing so, the scale of parameters gets increased. Currently, the correspondence between the complexity of the problem and the magnitude of the model parameters is still unavailable. However, it is believed that one factor contributing to the success of DL is the massive representation space which is formed by the large quantity of parameters. The second benefit is that a more powerful heterogeneous system could be constructed via stacking. As mentioned in [78], the controller of DNC is free to choose. For multi-layer LSTM, although the configuration of each layer can be different from each other, practice from implementation tends to put constraints on the flexibility and always requires homogenous layers. But the configurations and operations of DNCs can be independent of each other, which means the stacked DNCs in Fig. 5.2 is a more flexible heterogeneous system. In fact, the two DNCs stacked together are slightly different from one another in the application part. However, by cooperation harmonically they can form a more capable system than functioning alone. In addition, specialties of EEG data analysis demand modifications to the original DNC for a proper instantiation. The following subsections details the enhancement to the internal structures and operations of DNC in accordance with the paradigm of the EEG data analysis.

5.2.2.1 Computation Graph

For EEG signal processing, a series of EEG segments (input) are related or associated to a certain mind state or cognitive phenomenon (label) in general. By terming them input and label respectively to free from the neurological context, the problem formed in EEG domain can be delivered to the machine learning domain to solve. By treating EEG signals as time series data and model the problem via RNN, it is of a many-to-one mapping paradigm. The original DNC structure is of the many-to-many type with a dependency graph drawn in Fig. 5.3(A). Hence, the first step for applying to EEG is to deduce the formula for a corresponding many-to-one mapping.

Adopting the same symbols as in [78], the computation graph of the many-to-many mapping is governed by the following (5.1) - (5.4). Here x_t denotes the input from the dataset, or the sample itself at time step t. r_{t-1}^{i} denotes the output from i^{th} read head at last time step. They are concatenated together to form the total input to the system. h_t^l denotes the hidden states of the l^{th} hidden layer. Usually, the l^{th} hidden layer might directly have input from the input layer instead of the right below $(l-1)^{\text{th}}$ hidden layer if they are recurrent in different directions, for example, bi-directional RNN. W_y denotes the weights from the concatenated hidden outputs after flattening, to a fully-connected layer, which is viewed as part of the computational network. W_ξ denotes the collective parameters involving the memory network, all of its parameters are optimized during the end-to-end training process. A better illustration of these symbols is shown in Fig. 5.3.

$$\chi_t = [x_t; r_{t-1}^1; \cdots; r_{t-1}^R] \tag{5.1}$$

$$v_t = W_y[h_t^1; \cdots; h_t^L] \tag{5.2}$$

$$\xi_t = W_\xi[h_t^1; \cdots; h_t^L] \tag{5.3}$$

$$y_t = v_t + W_r[r_t^1; \cdots; r_t^R] \tag{5.4}$$

Other formulas such as the ones governing the update of the controller network are left to the next section; for instance, how to compute h_t^i from χ_t and h_t^{i-1}, etc. Fig. 5.3(A) is only a corresponding simplified version for demonstration.

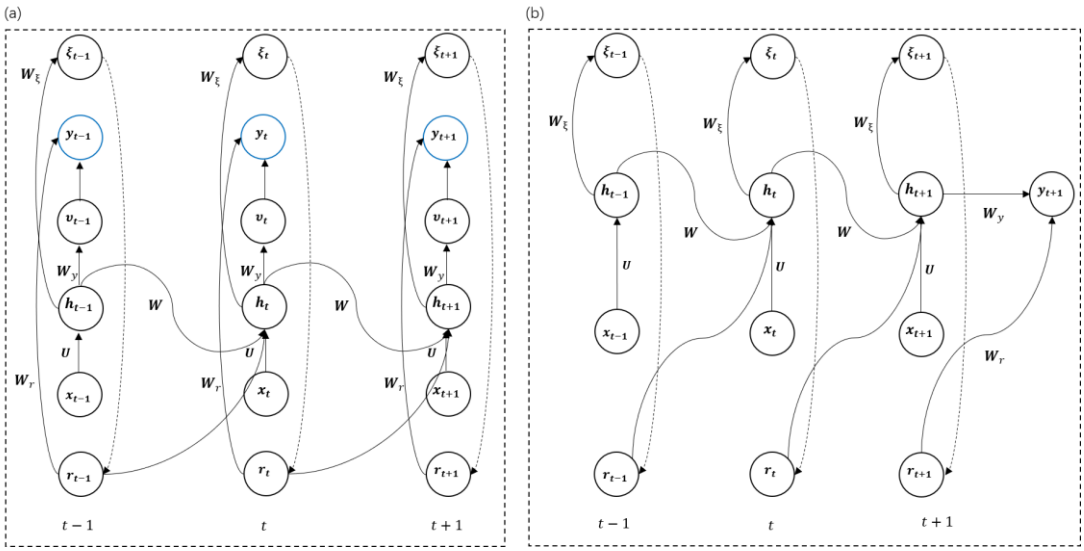


Fig. 5.3 (A) Computation graph dependency for the many-to-many paradigm of DNC; (B) Computation graph dependency for the many-to-one paradigm of DNC.

For many-to-one mapping, the learning signal or label only exists for the last time step; hence, v_t is only necessary for the last time step T. This reasoning leads to the following formulas (5.5-6) and the corresponding computation graph as in Fig. 5.3(B).

$$\xi_t = W_\xi[h_t^1; \cdots; h_t^L] \tag{5.5}$$

$$y_T = W_y[h_T^1; \cdots; h_T^L] + W_r[r_T^1; \cdots; r_T^R] \tag{5.6}$$

5.2.2.2 Controller Network

The next concern of designing stacked DNCs for EEG data analysis lies in the various EEG formats that the research community adopts. Due to the fact that success of the DNN mostly attributes to the utilization of convolutional operations to iteratively exploit the local structures of the input data, EEG topographical data are used here to further tailor the network structure.

However, the controller in [78] is constructed from a traditional LSTM network. It may fail to merit the spatial correlation of 2D inputs. Inspired by the work in [120], the structure of a recurrent convolutional network is put into consideration here. To effectively blend convolutional operation with recurrent operation from a computation efficiency perspective, a modified version of gated recurrent unit (GRU) governed by the following formulas is specially designed as the building block for controller network:

$$\chi_t = \text{concat}(I_t, S_{t-1}, M_{t-1}) \tag{5.7}$$

$$z_t = \sigma(\text{conv}(\chi_t, W_z)) \tag{5.8}$$

$$r_t = \sigma(\text{conv}(\chi_t, W_r) + 1.0) \tag{5.9}$$

$$c_t = \tanh(\text{conv}(\chi_t, W_i)) \tag{5.10}$$

$$h_t = h_{t-1} * r_t + c_t * z_t \tag{5.11}$$

In (5.7), I_t denotes the current input to the network. S_{t-1} is the state of the controller network and M_{t-1} is the related memory information at the previous time step. To address the similarity, it is preferable to compare between normalized vectors of which no component supremely dominates; so, all the activation functions chosen in

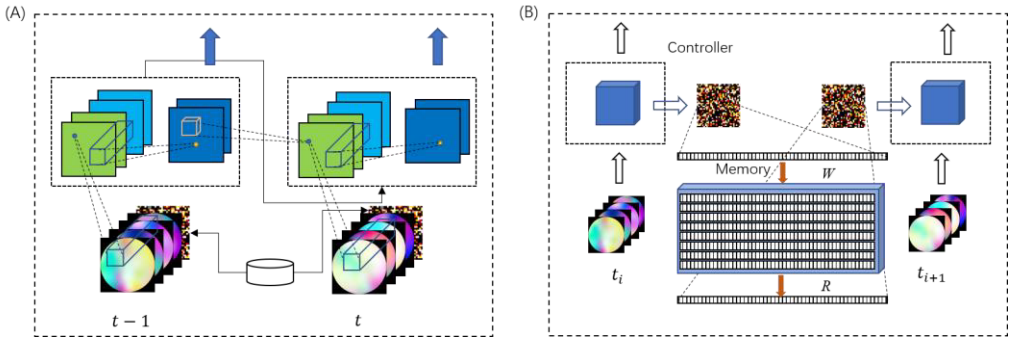


Fig. 5.4 (A) Controller network constituted by a recurrent convolutional network; Interactions with memory is substantially simplified just for indication (B) Special tailored controller of DNC: a recurrent convolutional structure substituted the original LSTM; Feature maps have to be flattened and concatenated into vectors to be stored into the external memory.

(5.7-10) are either sigmoid or hyper-tangent functions. The design of the controller network complying with the above formulas is demonstrated in Fig. 5.4(A). Overall, the controller network can be regarded as a convolutional gated recurrent cell network (Conv.GRU).

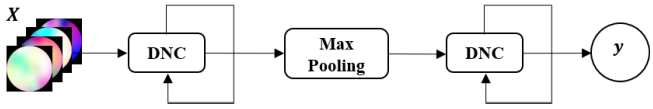


Fig. 5.5 The overall structure of stacked DNCs; Note the maximum pooling layer between consecutive DNCs.

5.2.2.3 External Memory

The last consideration of the proposed architecture lies in the memory layout. The external memory in [78] is in the matrix form, of which each row represents a context- or problem-dependent granular vector from the internal state perspective, while the number of rows represents the capacity. Working with EEG topographical data results in that all intermediate processing representations are essentially 2D. To cater for the requirement of storing the internal states into the memory in row vector form, a conversion between 1D and 2D layouts is specially designed for the whole structure to work as expected, as in Fig. 5.4(B). It is pointed out that it is not the optimal way for exploring the potential of stacked DNCs model, and more appropriate versions will be considered in future work.

The overall architecture of the network is recapped in Fig. 5.5. Here only two DNCs are stacked together but there is no constraint on the number of DNCs involved. A maximal pooling layer is also introduced between consecutive DNCs to reduce the dimension, which is a stereotypical treatment in DNNs. DNC is a general architecture, for instance, the size of external memory is still problem-depended; so, the details of the configurations are delayed to the experiment section. However, it is always the architecture as in Fig. 5.5 which is adopted to analyze different EEG datasets.

5.2.3 Experiments and Results

Two EEG datasets are used to justify the feasibility and elegance of the proposed stacked DNCs.

5.2.3.1 Mind Load EEG Dataset

To validate the feasibility of the proposed model, the EEG data captured from the working memory experiment in section 2.3.3 are utilized first.

As in Fig. 2.6, the duration from the display of the randomly generated letter set to the response of the test subjects is called a trial. The EEG data lasting from $t = 1$ to $t = 3500$ ms is taken as input data, which is segmented into 7 non-overlapping consecutive pieces with each equal to 500 ms length. For each segment, Fast Fourier Transform (FFT) is performed on the time series data to calculate the power spectrum of the signal. The average power for EEG sub-bands, i.e., theta, alpha and beta, are put into consideration here. The same approach is taken as in [71] to map the spectrum components into topographies. The transform procedure is indicated in Fig. 2.1. The size of test set, i.e., number of letters, is treated as mind load label.

Table 5.1 Configurations of stacked DNCs for mind load EEG data analysis

	Controller		**Memory**	
DNC1	Structure	RNN	Slot Size	16
	Cell Type	Conv. GRU	Word Size	1024
	Filter Size	5x5	Read Head	4
	Feature Map	12	Write Head	1
DNC2	Structure	RNN	Slot Size	16
	Cell Type	Conv. GRU	Word Size	256
	Filter Size	5x5	Read Head	4
	Feature Map	12	Write Head	1

To benchmark different models for evaluating the performance, an identical data preparation method is taken as in [71]. First, one subject is excluded for each fold, and the trials of all left subjects are combined and shuffled. Then the number of trials equal to the excluded subject in the current fold are picked out for validation and all the remains are used for training. This leave-one-subject-out process is repeated for all the test subjects.

The model is trained with the configuration as in Table 5.1. The batch number is set to 10 percent of training samples, i.e., around 225. The training converges after 5 epochs but the whole process lasts for 40 epochs. The early stop strategy is taken as in [71] to prevent overfitting. The statistics for testing and comparison are in Table 5.2. In Table 5.2, the first three methods are all from [71]. They are convolutional network, recurrent network with LSTM cell and recurrent-convolutional network respectively. "Mix" means features extracted by conventional network are fed into recurrent network for final inference. It also gains the name of recurrent-conventional network, which is of better performance over other methods in general.

Although the results are still subject-depended and not the state-of-the-art results for all folds, it is eminent that the stacked DNCs achieve the best average accuracy compared with other methods. It is believed that the less optimal results for some folds are due to some factors such as only a portion of parameter space is explored when seeking the optimal ones. With more computation resource and effort, there is still some margin of which the results could be improved.

Table 5.2 Mind state classification accuracies for each fold and means*

Test Subject	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**	**S11**	**S12**	**S13**	**Mean**
1D-Conv	88.3	72.5	93.9	97.5	98.3	98	98.2	100	98.5	94.5	88.5	79.5	45.9	**88.7**
LSTM	56.7	73.5	92.2	99	99.4	99.5	98.9	100	100	97.7	99	88	59.1	**89.5**
Mix	88.9	76.5	93.3	99	100	98	100	98.5	99	96.8	96.5	91	46.8	**91.1**
S-DNC	82.2	75.9	91.5	100	99.5	99.5	98.4	100	98.6	97.3	100	90.4	68.2	**92.4**

*part of the results is directly from [71]

5.2.3.2 Driving Task EEG Dataset

This experiment makes use of the EEG dataset captured from the driving experiment described in section 2.3.1.

For data pre-processing, the procedures resemble the treatment as in [121]. However, in [121], it works with graphs as shown in Fig. 2.1(B), after manually adjusting the channel sequence for x-axis. As mentioned above, in order to utilize neural networks to automatically explore the spatial relations between channel locations, it is preferred to work with topographies as in Fig. 2.1(C). Because the mapping from Fig. 2.1(B) to (C) requires interpolation and extrapolation of the original channel values, all the channels are retained here to ensure the interpolating or extrapolating accuracies. But it is still common to select certain channels based on the experience and knowledge to analyze as in [121]. The less picky on EEG data means more information can be kept during the transformation. The topographical data preparation procedures are identical to the previous experiment.

Table 5.3 Configurations of stacked DNCs for driving EEG data analysis

	Controller		**Memory**	
DNC1	Structure	RNN	Slot Size	32
	Cell Type	Conv. GRU	Word Size	256
	Filter Size	5x5	Read Head	2
	Feature Map	12	Write Head	1
DNC2	Structure	RNN	Slot Size	32
	Cell Type	Conv. GRU	Word Size	256
	Filter Size	5x5	Read Head	2
	Feature Map	12	Write Head	1

For simplicity, the dataset is depicted abstractly in a machine learning context. The pre-processed multi-channel waveform EEG data captured during the baseline period (600 ms immediate before deviation onset) after transformation is denoted by X. The corresponding RT that is used to measure the drowsiness of test subjects is denoted by y. Each trial corresponds to a sample pair (X, y). The problem is treated as a regression to predict RT or y given baseline data X.

EEG data of 5 subjects with the leave-one-subject-out method is used for training and test. With the configuration as in Table 5.3, the model takes around 2.5% of the total training samples as batch number trained for 500 iterations. The initial learning rate is 0.0001 and decays by a factor 0.8 for every 100 iterations. The model is evaluated 5 times for each subject and the averaged root-mean-square errors (RMSEs) of the predicted RTs are treated as the final performance indicators.

Table 5.4 Comparison of RMSE by different models

Model	**Indicator**	**S1**	**S2**	**S3**	**S4**	**S5**	**MEAN**
1D-Conv	Accuracy	0.369	0.763	0.598	0.557	0.732	**0.604**
	Corr. Coef.	0.322	0.483	0.610	0.435	0.083	0.387
LSTM	Accuracy	0.344	0.943	0.630	0.390	0.733	**0.608**
	Corr. Coef.	0.434	0.443	0.553	0.462	0.158	0.410
Mix	Accuracy	0.383	0.779	0.599	0.552	0.740	**0.611**
	Corr. Coef.	0.324	0.479	0.600	0.432	0.075	0.382
S-DNC	Accuracy	0.380	0.769	0.578	0.485	0.717	**0.586**
	Corr. Coef.	0.339	0.413	0.472	0.389	0.138	0.350

For comparison, all the models in [71] have been re-implemented as benchmark methods. The statistics are shown in Table 5.4. From Table 5.4, it is manifest that the performance achieved by the proposed stacked DNCs is the best among all. The correlation coefficients between various models are also compared. The overall comparable coefficients demonstrate the practicability of the proposed model. It is believed that by continuing to fine-tune the parameters, there is still room to reduce the averaged RMSE. Even though, the potential of utilizing the general stacked DNC for EEG signal analysis is well demonstrated.

5.2.4 Discussion and Future Work

Neural networks are criticized as black-boxes despite their excellent performance. It is helpful to investigate the behaviour of the network to better understand the underlying working principles or to inspire the design of new architectures. In the following, the EEG dataset captured during the driving task is focused to carry on the discussion.

From the design principles and test setup of the driving experiment, it is obvious that visuomotor sensory and mind state such as drowsiness, are key to the subjects' cognitive process during the experiments. For visuomotor sensory, due to brain structure [4], it is well-known that the occipital part is the most active area [122, 123]. And it is also believed that the frontal cortex is correlated with the drowsiness/alertness and attention variance [124, 125]. Therefore, it is inclined to expect that such priori could be reflected in the trained model.

To inspect the behaviour of the stacked DNCs after training with the experimental data, the trained parameters is restored first. Then a series of white noise topographical images are input into the network. By observing the manipulated feature maps, it is to seek whether there exist such correlations with neuroscience discoveries or not. The reason for choosing white noise images is that it is intended to know whether these original average-distributed power values can be enhanced or saturated by the network especially for some local regions. The feature maps to be investigated are drawn from the first DNC, because it is believed that convolutional layers in lower hierarchy of neural networks tends to extract concrete features [3], which is easier for illustration. Four feature maps selected from a total 12 are drawn in Fig. 5.6 for demonstration and comparison.

It is interesting to note that the feature map 1 and 2 mainly focus on the occipital area, on conjecture that signals related to visuomotor should be specifically captured by the network model for better performance [122, 123]. For feature map 3 and 4, they are concentrating on the contralateral frontal cortices. Because the recruited test subjects are all righthanded and mind state is highly demanded in the whole process, it makes sense and shows coincidence with the conclusion in [124, 125]. In general, all these parts that got modulated are in accordance with the expectations.

Based on the above observation or discussion, it can be concluded that besides asynchronization of recurrence, DNC is also endowed with some attention or local concentration mechanism. Complexity is usually an unavoidable topic for EEG data, plus the limited understanding of cognitive processes, hence it tends to be a hard problem to manually extract features from EEG data to help interpret the functionalities of the brain for a specific experimental setup. However, such a local concentration mechanism of the proposed model can effectively select the targeted regions in favour of the optimal result, which in turn unveils which area of the brain is more involved in

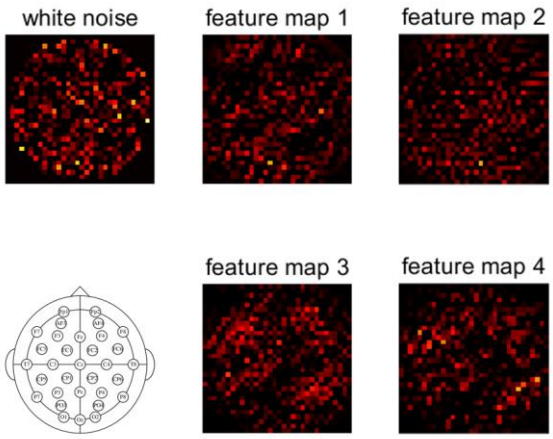


Fig. 5.6 The white noise input and the corresponding feature maps. The positional relation in each figure is indicated as in the atlas figure in the bottom left corner.

a specific task than others. Put it together, all these characteristics of DNC especially the stacked version reveal the feasibility and promising potentiality to be utilized for EEG data analysis.

Another aspect worthwhile to mention is the way that memory gets used. As aforementioned, the modification to the memory-related structure is plausible but may not the optimal approach. One difficulty lies in the implementation of the DNC itself. The original implementation is by analogous to the modern computer architecture. Although DNC has been successfully applied to solve various complicated problems, following the architecture of modern computers especially the memory part requires two challenges to be solved, i.e., how to decide the address for accessing data and how to evaluate the freeness of memory locations. In fact, several RNNs internally to DNC architecture are designed to cater for these challenges, but all these avoidably add to the complexity of the whole architecture and make the adjustment or modification to the original memory part extremely difficult. These problems will be considered in future work to seek some ingenious and direct way to reduce the complexity.

5.3 Memory Augmented Convolutional Network

DNC tries to establish an enriched computation model by integrating a large memory, and two extra LSTM networks are tailored for the memory mechanism. As mentioned in the last section, following the architecture of modern computer avoidably adds the complexity to the whole architecture and make the adjustment and extension to the memory module extremely difficult. However, it does not indicate in the negative way, it is because of the limited understanding of human memory system, the inspiration from it is not that intuitive and difficult to manage. In fact, leading companies such as Google and Facebook are still working on more sophisticated neural networks integrating memory to solve more difficult tasks.

However, certain DL applications are not targeting sequential data, such as image classification, face recognition, cancer prediction, to name a few. These applications are based on the variations in CNN architectures, which are generally regarded as computation-oriented. This viewpoint is more obvious if the neural network is treated as a functional approximator from the arithmetic perspective. Although during the training process, the network interacts with all the training samples, it is still recognized that it is mainly for learning optimal filters for the targeted problem instead

of learning to memorize. Hence, the impressive performance of CNN relies on the features being effectively extracted from input data via these optimal filters [126]. Alternatively, it means that the prediction is mainly based on the computation carried out on the current input sample, and the relations with historical samples are very vague.

But for a group of images containing the same object, they share many features common to that class of object. For CNN, categorization of the content in a specific image mainly harnesses the knowledge retrieved from the current image via applying learned filters. Instead, humans tend to make the judgement comprehensively via cognitive process such as familiarity or even recall if necessary, all are involving the memory system, especially recall [127]. The above consideration encourages to devise a memory module to couple with CNN for boosting the performance.

In fact, the idea of utilizing memory as part of computation has its own history. Kant proposed the concepts of memory elements and the interconnections between the memory elements and their functionalities in [111]. John Hopfield introduced Hopfield network, working in a content-addressable or associative way with binary threshold nodes in [128]. In this section, self-organizing map (SOM) invented by Finnish professor Teuvo Kohonen as the memory module is proposed to be used for fertilizing computation [129]. SOM is based on the biological models of neural system [130, 131], and it can capture the intuitive essence of memory system and meanwhile maintain its simplicity.

5.3.1 SOM Recap

SOM used to be a popular neural network model and could find its various applications in numerous fields [132]. It belongs to the category of competitive learning networks and implements the winner-take-all learning strategy [129]. The arrangement of neurons plus the unsupervised learning paradigm enable it suitable to intrinsically represent the features to the input sample space in a topological way automatically. The updating process, which drives the best matching unit (BMU) and its neighbouring neurons towards the value of new input sample, endows SOM the potential of tracing all past sample values, if the lattice of SOM is large enough. Nevertheless, averaging between samples belonging to the same category and differentiating of samples belonging to the different categories are critical for learning. It can encourage the features common to all samples to emerge, which is critical for model performance. As mentioned above, for CNN, where the learned filters are acting on the input data for final classification or regression, the process can be considered computation oriented. Instead for SOM, the weights are aggregating on input data for feature abstraction, it can be considered memory oriented at least to some extent. Such a perspective makes it opt for memory network such as in LAMSTAR [114]. Other utilizations of SOM such as visualization of sample space can be found in [133, 134].

To facilitate the understanding of the designed network and its configurations in the following sections, the training process of SOM is particularly addressed below. It also highlights the aspects which can be traded off for better performance, such as learning rate manipulation, neighbourhood calculation, etc.

1. Randomly initialize the weights of neurons in the lattice, which is of dimensions $H \times W$. H denotes the height, and W denotes the width of the lattice.
2. Choose an input sample from the training set and present to the lattice. It can be in a batched way.

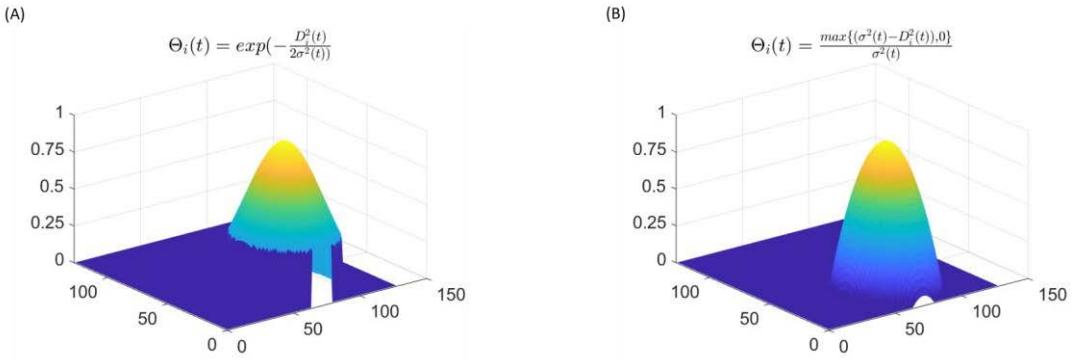


Fig. 5.7 Learning rates modulated by different neighborhood functions

3. Calculate the Euclidean distances between the input sample $\vec{v} = (v_1, v_2, \cdots, v_n)$ and weight $\vec{w} = (w_1, w_2, \cdots, w_n)$ of each neuron, with metric such as $\|\vec{v} - \vec{w}\|_2 = \sqrt{\sum_{i=1}^{n}(v_i - w_i)^2}$. The node with the least distance is selected and named BMU.
4. Decide the neighbourhood of BMU for the current step.

   The neighbourhood is usually a disc mask, with initial radius σ_0, and shrinks according to formula such as:

$$\sigma(t) = \sigma_0 \exp(-t/\lambda) \tag{5.12}$$

   λ is regarded as the time constant, in the form such as $\lambda = T/\ln \sigma_0$, with T the total number of learning steps or training iterations.

   A boundary condition analysis reveals that

$$\sigma(T) = \sigma_0 \exp(-T/(T/\ln \sigma_0)) = 1 \tag{5.13}$$

   It is with the expectation that the last step update only confines to the BMU itself.
5. Adjust the weight vectors for the neighbouring neurons.

   Determining the learning rate for the current step according to (5.14):

$$\eta(t) = \eta_0 \exp(-t/\lambda) \tag{5.14}$$

   Calculate the Euclidean distance D_i between the BMU and neuron i based on their locations in the lattice, and compute the corresponding distance influence factor $\Theta_i(t)$ according to (5.15):

$$\Theta_i(t) = \exp\left(-D_i^2/2\sigma^2(t)\right) \tag{5.15}$$

   The adjustment for neuron i is with accordance of the formula

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \Theta(t)\eta(t)\left(\vec{v}(t) - \vec{w}_i(t)\right) \tag{5.16}$$

6. Repeat 2-5 for T steps.

Notably, in the above procedures, as long as certain paradigms can be satisfied, most formulas are not unique but can have alternatives. For example, the radius of the current BMU neighbourhood can also be decided as in (5.17):

$$\sigma(t) = \sigma_0 - (\sigma_0 - 1)\, t/T \tag{5.17}$$

$\eta(t)$ can be alternatively updated as in (5.18):

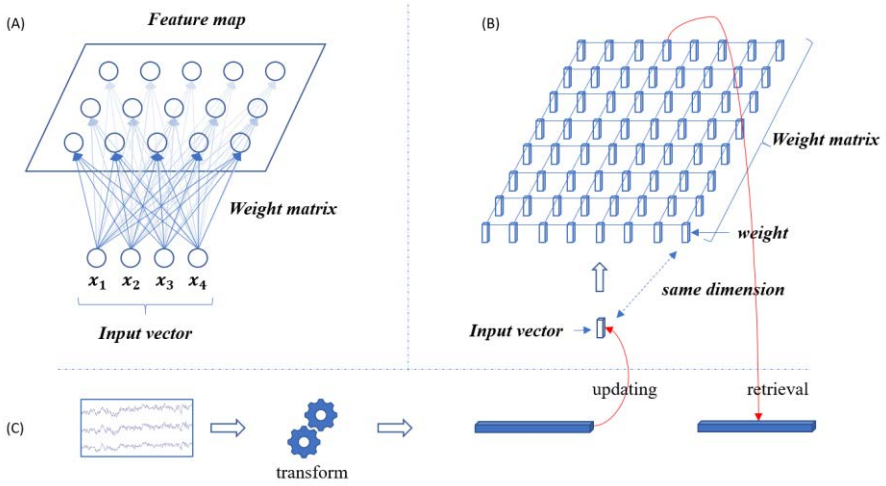


Fig. 5.8 SOM structure (A) Illustration from conceptual perspective; (B) Illustration from pragmatic perspective (C) Dimension reduction of EEG data and its interaction with SOM module.

$$\eta(t) = \eta_0(1 - (t-1)/T) \tag{5.18}$$

Another method for amending $\Theta(t)$ is according to (5.19):

$$\Theta(t) = \max\{(\sigma^2(t) - D_i^2), 0\}/\sigma^2(t) \tag{5.19}$$

Fig. 5.7 shows different neighborhood modulated learning rates for a given step. Note in Fig. 5.7(A), there is a truncation beyond the current radius. The choice between these formulas inevitably increase the parameter space for exploration for fine-tuning. But it provides the plausibility to adopt different formulas for different problems in achieving the best performance. However, to provide a rigid guideline for the selection of these behavioural paradigms, which affect the dynamics of SOM during learning, is still a future work.

5.3.2 Network Architecture

In this part, a hybrid architecture that jointly combines computation and memory by incorporating a SOM into the CNN is designed. The general structure of SOM is shown in Fig. 5.8(A). Each component of the input vector connects to a neuron which is arranged in a two-dimensional lattice. Assume the dimension of input sample is N, let i denote the ith input component and j denote the jth neuron, then the corresponding weights are $[w_{i_1}^j, w_{i_2}^j, \cdots, w_{i_N}^j]$.

To conceive a memory module based on SOM, a change in view is needed. As in Fig. 5.8(B), each neuron can be thought of storing a vector that has the same dimension as the input sample. And the weights linking input and neurons are always of fixed value 1. In fact, at the initial stage of neural network development, the plasticity of synapse for learning is not well understood. Therefore, the weights are always set fixed, and only neurons modulate the relied signals [28]. However, to consider from the viewpoint Fig. 5.8(B) can ease the conceivable barrier when adopting SOM as memory part of the network. Viewing the structure of SOM in either Fig. 5.8(A) or Fig. 5.8(B) takes the same learning or updating mechanism.

The next consideration is the content to be stored in the neurons. In applications, the data samples are usually of high dimensions, and to directly store the raw data is inapplicable. An alternative is to store some transformed features of lower dimensions, as shown in Fig. 5.8(C). The criterion is that the transform should be capable of

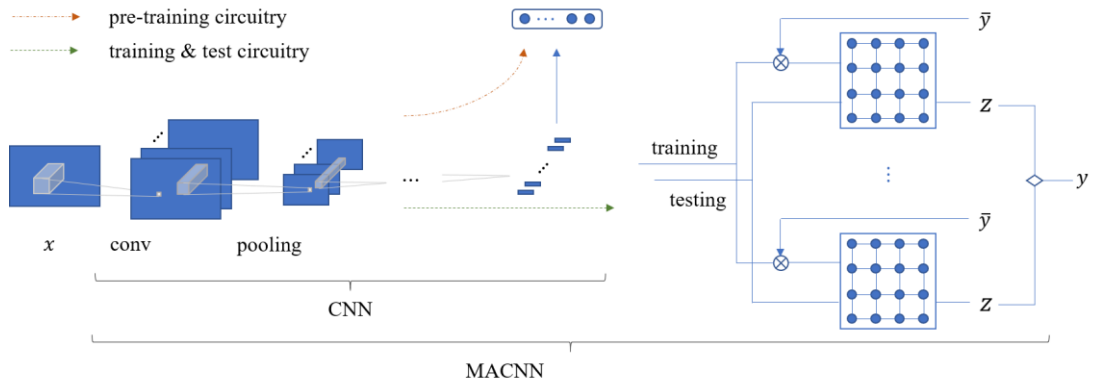


Fig. 5.9 The overall network architecture.

extracting suitable features from the original data. Based on these reflections, the designed neural network architecture is shown in Fig. 5.9.

In Fig. 5.9, the overall architecture is a combination of CNN and SOM, abbreviated as MACNN (memory augmented convolutional neural network). Only classification problem is considered in this work, and there is a SOM block accompanying each class. The memory or SOM module consists of individual SOM blocks. To articulate the original input into features for memorization and refinement in SOM blocks, a CNN is introduced to act as the transform, based on the idea of transfer learning. To have the CNN suitable for extracting features, a pre-training stage (or circuitry) is introduced. The pretraining stage is same as the general application of CNN. In detail, the CNN is trained first, and then the weights are fixed and transferred into the second stage (training) to extract the features for further manipulation. Subsequently, the extracted features and SOM are interacted for the final computation goal, namely, classification.

Depending on whether it is the training stage or test stage, the inputs are routed into the SOM in different ways. For training, the input (extracted feature) is gated by the ground-truth label. Assuming there are N classes, let x denote the input to the SOM module, and $g(x) = \{g_1(x), g_2(x), \cdots, g_N(x)\}$ denote the gate function, where $g_i(x) = 1$ if $x \in C_i$ and $g_i(x) = N/A$ (means unavailable) otherwise. The input to i^{th} SOM module is $x \cdot g_i(x)$. For example, suppose the current input belongs to the first class C_1. Then, the input goes only to the first SOM block and update the block according to the SOM learning method. All other SOM blocks are inactive for the current input. For the test stage, the input goes to every SOM block for computation.

The motivation for the design of the network architecture in Fig. 5.9 draws inspirations from the following biological facts closely related to neuroscience. First, the memory system is a distributed system, and knowledge are stored in different areas of neocortex [135]. Second, long-term memory induces and depends on the modulation of synapse plasticity [136]. For a classification problem, different categories usually represent different concepts. The intention for incorporating separative SOM blocks is to cater for the needs of representing different categories in a distributive manner. The gating mechanism is devised to solve the difficulty of learning process. SOM favours the fact that learning roots in synapse plasticity alteration. If the input is routed into all SOM blocks during training, the one corresponding to the correct label shall update the weights (or content stored in neurons) towards the favouring direction. The rest should update in some adversarial directions. The problem is in high dimensional space, it is difficult to define the adversarial direction. Therefore, an alternative is to let all SOM

blocks update only towards the favouring direction. It means, only the SOM block corresponding to the correct label learns per input sample, with others remaining inactive. Other methods for training might exist, but this method is simple and computationally efficient.

For testing, all weights involved in the whole architecture are frozen. The input is routed into every SOM block directly to find the BMU. Then the same metric (usually the Euclidean distance) between the input and all BMU candidates are calculated, among which the minimum one is located. The index of the minimum distance is regarded as the corresponding category to which the input belongs.

The postulation underpinning the above procedure is that during updating, the SOM block corresponding to a certain category to which some samples belong, is always learning from these samples. Different features, which are supposed to be captured by different SOM blocks, differentiate samples from each other. These learned features are not from one sample, but aggregating from all the historical samples and get memorized by the SOM. This process resembles the memory forming procedure, aka perception, consolidation, etc. And compared with the vanilla CNN, it is believed this memory-incorporating model can exhibit better performance.

5.3.3 Experiment and Result

The EEG dataset used in this experiment is captured from the experiment described in section 2.3.2. For most cases, EEG signals captured during certain physiological or cognitive process are serial data, which can be analyzed by constructing RNN with LSTM cell. But it is not always the case, for instance, the P300 phenomenon in oddball tasks such as the experiment in section 2.3.2. The imminent EEG amplitude time-locked to the appearance of the visual stimulus (target) is in a rather limited effective duration (0.7 second at the maximum), and RNN is difficult to apply in this circumstance. Additionally, EEG signals are complicated bio-signals due to high intra-subject and cross-subject variance and the low signal to noise ratio (SNR). Those factors render sole application of CNN in this situation not as fruitful as in other cases. Hence, the proposed model is very appealing for this case.

For pre-processing, the raw EEG data are first rectified via the PREP pipeline [137], and then subjected to multiple artifact rejection algorithm (MARA) for ICA-based artifacts removal [138]. Next, the signals are down-sampled to 256 Hz, following segmentation into [0, 1.0] second intervals time locked to the stimulus onset. To mitigate the effect of outliers, epochs with incorrect button presses are removed. Then, the mean baseline is removed from each channel in each epoch. This leads to a final 382 target trials and 3249 non-target trials.

The placement of 64 channels is complied with the international 10-20 system. Because VEP originates from the occipital cortex, which is dominantly involved in receiving and interpreting visual signals, signals from 12 channels, which mainly cover the parietal and occipital areas are selected for analysis. Because samples for these two classes (target vs non-target) is of an imbalance case, the samples are not blended then divided according to a certain ratio for training, validation and testing. Considering the cross-subject variance of the EEG data, a more tactic treatment is adopted. For training, to prevent the misbehaviour during the training process of the proposed network, non-target samples are resampled to match the target samples. For example, suppose the batch number is N; $N/2$ non-target samples are resampled to blend with target samples for each training iteration.

Table 5.5 Network Configuration for EEG Dataset Experiment

Network (CNN)					Stages		
No.	Layer	#Filter	Kernel	Act.	Pre-training	Training	Test
1	Conv2D	16	1x5	tanh	x*	x	x
	AvgPool	-	1x2	-	x	x	x
2	Conv2D	32	1x5	tanh	x	x	x
	AvgPool	-	1x2	-	x	x	x
3	FC	32	-	tanh	x	x	x
4	FC	2	-	Softmax	x		
Network (SOM)							
0	Height	Width	Depth	Radius		x	x
	128	128	32	8			
1	Height	Width	Depth	Radius		x	x
	32	32	32	8			

*x indicates the presence of the corresponding layer

For the network configuration in Table 5.5, the CNN part is straightforward. Notably, non-target samples with labels 0 dominate over target samples with label 1. To cater to the appropriate learned knowledge representation, the dimensions of separate SOM blocks are more diverse. The dimensions of SOM block 0 are quadruple as those of block 1. Furthermore, although the change of viewpoint from Fig. 5.10(A) to Fig. 5.10(B) makes the original EEG data no eminent difference from image data, the multi-channel property still makes the convolutions along the height under dispute. Thus, the convolution is 1D and only along the width dimension. The training procedure is identical to the previous experiments. Because EEG data tend to display low SNR, the learning rate of SOM is set to 0.01 to cater for the procedural refining process of SOM update.

For test, the data from the last four subjects are reserved to perform the leave-subjects-out test, which is more objective to assess model performance. To objectively assess the performance of the proposed network architecture, the whole procedures has been repeated for 5 times. Because imperfect data can lead to high variance predictions, in practice usually the best model is chosen for inference. As for performance indicator, maximum metric of multiple runs, either accuracy or F1 score is considered. For each pre-training CNN, the transferred weights of CNN are used for 3 times to train MACNN, and the maximal test accuracy is taken as the performance indicator for the current fold. The final comparison is between the maximal prediction accuracy of CNN and MACNN respectively.

With the above descriptions, taking a learning rate of 0.001 with batch number 64, the model is pre-trained (or train the sole CNN) for 10,000 iterations. To stabilize the pre-training process, the learning rate is decayed by a factor of 0.96 for every 100 iterations. After pretraining, with a learning rate 0.01 for SOM module, the MACNN is trained for 10,000 iterations as well. Part of test data are used to monitor both the pre-

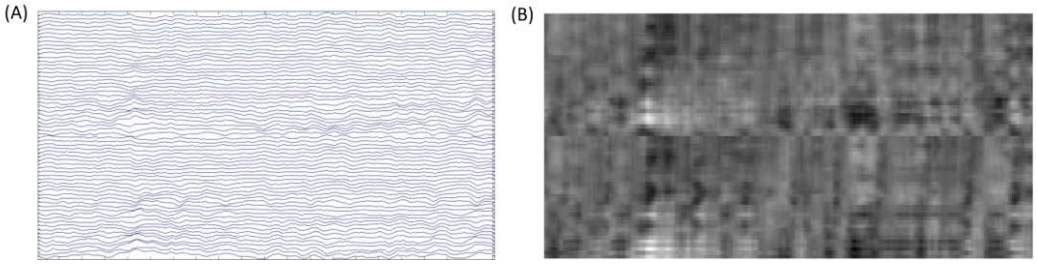


Fig. 5.10 Viewpoints of EEG from different perspectives. (A) Multi-channel data; (B) Image data.

Table 5.6 Test Statistics for MNIST Experiment

		CNN (↓)	MACNN (→)			
No.			1	2	3	Maximum
1	Accuracy	0.7884	0.7858	0.7720	0.7934	**0.7934**
	F1 Score	0.4509	0.4516	0.4290	0.4569	**0.4569**
2	Accuracy	**0.8702**	0.8551	0.8564	0.8589	0.8589
	F1 Score	0.4550	0.4549	0.4818	0.4766	**0.4818**
3	Accuracy	**0.8513**	0.8387	0.8224	0.8236	0.8387
	F1 Score	0.4271	0.4285	0.4291	0.4308	**0.4308**
4	Accuracy	**0.8488**	0.8274	0.8186	0.8136	0.8274
	F1 Score	**0.4230**	0.4170	0.3999	0.3983	0.4170
5	Accuracy	**0.8299**	0.8148	0.8085	0.8047	0.8148
	F1 Score	0.3349	0.3466	0.3448	0.3459	**0.3466**
Maximum	Accuracy	**0.8702**	-	-	-	0.8589
	F1 Score	0.4550	-	-	-	**0.4818**

training and training processes and no overfitting is observed. After training of both CNN and MACNN, the statistics for the whole test set are shown in Table 5.6.

As from Table 5.6, it is obvious CNN achieves the best accuracy among multiple runs. However, accuracy is inappropriate for imbalance data because the training process can drive the model to misclassify the minority and still retain high accuracy. That is the reason for training CNN, resampling is adopted to prevent this perversion. And in this circumstance, F1 score is usually taken as the performance indicator [139]. Although MACNN fails in beating CNN on accuracy, however, it is supreme on F1 score as in Table 5.6, which indicates a promising application for imbalance data.

More comparison is made with results from [140] which use the same EEG dataset in studying different sampling strategies for imbalance data. The best result of the proposed MACNN in Table 5.7 suggests that by combining SOM and CNN together, harsh situations such as imbalance data can be effectively tackled.

Table 5.7 Test Statistics for MNIST Experiment

Model	**NN***	**SVM**	**CNN**	**MACNN**
Accuracy	0.694	0.756	0.8702	0.8589
F1 Score	0.309	0.288	0.4550	**0.4818**

*Multi-layer Perceptron

The distributions of features of all test samples extracted by CNN and MACNN separately are also illustrated in Fig. 5.11. It helps to understand the difficulty of highly accurate classification for both models and then try to understand MACNN's behavior. Fig. 5.11(A) illustrates the features from CNN via t-SNE, which demonstrates the separability of target and non-target samples. Even after several neural network operations such as convolution and pooling, the interleaving of features is not effectively resolved into a clear condition. In contrast, in Fig. 5.11(B), the features are organized in a more categorial way, approximating the intended dichotomic phenomena of the oddball task. However, the intrinsic non-separability of the problem still prohibits to produce a good prediction; nevertheless, it does not obscure the situation in Fig. 5.11(B) is better than that in Fig. 5.11(A).

5.3.4 Discussion and Future Work

The significance of human memory in the cognitive process which shapes people's daily life is self-evident. It is well-known that the human memory system

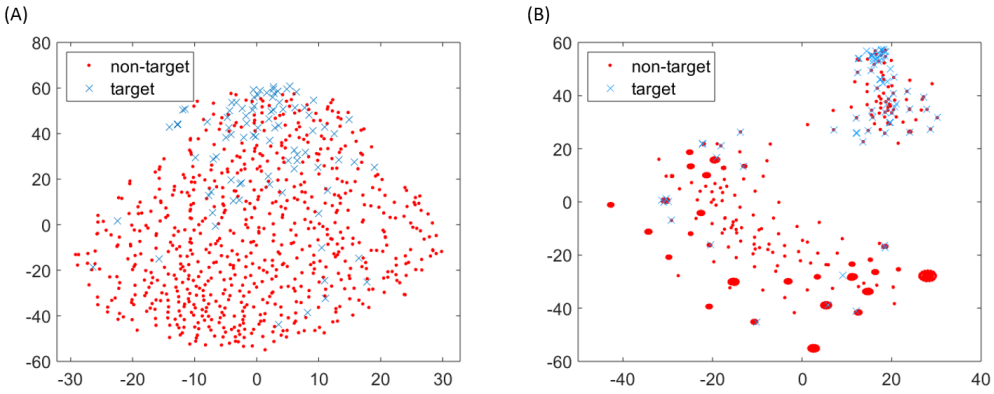


Fig. 5.11 Illustration of features of MNIST test samples from CNN and MACNN respectively t-SNE (A) CNN; (B) MACNN.

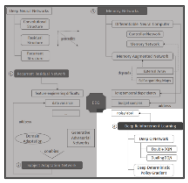
consists of working memory, short-term memory and long-term memory. Long-term memory is further categorized into explicit memory and implicit memory. How to inspire from human memory to design a more intelligent system is admittedly a meaningful research in this DL era, just as the work that was done in [78].

In this work, the proposed SOM module, which plays the memory role in CNN, is quite different from LSTM, at least from the brain-inspired computation perspective. LSTM mainly explores the temporal correlation during recursion when processing the current sample, and the content expires for new input. On the other hand, training an LSTM network is not for storing information, but rather to tune the weights associated with the input gate, forget gate and output gate. Generally, LSTM resembles the working memory.

However, SOM does store information of the samples. A boundary condition analysis can be performed to reveal that in an extreme case, the neighbourhood of BMU is just the unit itself as depicted previously. Furthermore, if the learning rate is 1, the BMU can be updated to the current sample. Given a sufficiently large lattice, all the input samples could be logged; nevertheless, this is not the way that SOM is intended to be used. Actually, if the current available samples (no matter training, validation or test) are sampled from a given space, the updating process of SOM can be treated as an interpolation and/or extrapolation of that space, meanwhile, smoothing among the samples. As in Fig. 5.8(B), each node in the lattice can be viewed as an instantiation of a class with some learned variance, which confers a better generalization as in the experiment.

The proposed network relies on a proper transform to retrieve the appropriate features to be stored in the SOM module. The current implementation is based on CNN, which is separately pre-trained. In future work, the opportunity for directly applying an end-to-end training will be investigated. Another aspect for improving the proposed work is the consideration of learning in an adversarial direction simultaneously during training. Currently, the SOM blocks always update weights in the direction that favours the correct label and neglect all negative samples. How to draw inspiration from the biological perspective and to architect a network capable of doing so is still under investigation.

6. Reinforcement Learning Paradigm



⌖ This chapter covers the bottom-right region of the research map in Fig. 2.8. By conforming to the methodology of reinforcement learning (RL), this chapter formulates a specific EEG data analysis task into an RL problem and solves it with modern DQN techniques, to address the noisy label issue and to enhance the performance of model generalization.

6.1 Background

One notorious problem for EEG research especially from the cognitive perspective via machine learning is the label problem. To exemplify this, considering the experiment which concerns driving safety research in section 2.3.1. Essentially, the target concept of this experiment is drowsiness. However, current limited achievements of neuroscience still prohibit direct probing of mind state, which means that there lacks direct measurement on the extent of drowsiness. Therefore, RT articulated in section 2.3.1 can mitigate this difficulty, by representing drowsiness on a specific time indirectly. However, RT can only be introduced occasionally during the experiment to introduce as minimal impact on the subjects as possible. Hence, the task is re-programmed as prediction of RT as frequency as possible to cover the whole experiment session, to reflect the drowsiness in a fine-grained manner.

However, indirect ways such as RT tend to be bias buried, which is more obvious in contrast to the image recognition task. Usually, there is no ambiguity for what is contained in an image. An image can be asserted that it either contains a dog or a cat, but not something in between. But due to factors such as a sudden drift of mind, different muscle motor trajectories, the measured RT is ineluctably noise contaminated, but these RTs are still presented as ground-truth labels. Retrospect of research during past years helps realize that following the supervised learning paradigm to train a complicated model in solely predicting RT chasing SOTA accuracy dooms to poor generalization during deployment. And a promising direction is to harness the available but inaccurate RT to trace the procedural change of mind state globally.

Besides the drawbacks induced by label problem mentioned above, for supervised learning including DNN, there exist other shortcomings, such as inefficiency data utilization and impracticability. Supervised learning models require the regularized input data to be paired with ground-truth labels for training and testing. If there is no label information, the corresponding EEG segments will be unavoidably discarded. Second, some models that require measured information such as RT for performance tuning only work in an offline manner and are unsuitable or inapplicable for online application, especially from an engineering perspective.

To mitigate these challenges, this part considers deep reinforcement learning (DRL), specifically deep Q-learning for analyzing EEG data. One eminent characteristic of reinforcement learning is the future reward can be used to assess the current action to take given the current state. In theory, a function involving the measured RT in some latency can be designed to yield the reward, based on which actions against the current state (EEG data segment) can be assessed. Another virtue is that for reinforcement learning, although the training phase requires the costly label

information (RT) obtained for reward calculation, for the testing stage, an optimal policy is always assumed and exploited for action selection. It means that no further deliberate articulated measured information is needed. In addition, EEG captured from the driving experiment is process-oriented, instead of event-oriented, which makes it suitable for this new computation paradigm, i.e., reinforcement learning. Furthermore, it is well-known that fatigue is the vital factor contributed to traffic accident fatal [141], and driving safety research via EEG targeting a portable fatigue detection system is of great practical values. These facts suggest a promising potential for analyzing the EEG signals of this BCI scenarios for real applications.

6.2 Deep Q-Network

6.2.1 Problem Formulation

The driving experiment is detailed in section 2.3.1. However, extra points are highlighted to bear the special considerations for applying deep Q-learning. First, the number of times for subjects participating in the experiment is not mandatory in order to encourage voluntary dedication although all subjects are paid. To ensure data quality, only one subject is hosted doing the experiment in one day. The number of sessions for each subject depends the number of his/her participations. Second, although there lacks an explicit formula linking RT with the mind state, it is postulated that they are positively correlated. The correlation can be estimated from the corresponding EEG data captured during the experiment. By introducing a reasonable threshold, RT is sufficient from the application perspective. For example, Any RT value above the threshold can be regarded as a poor control of the car, and a warning should be raised to prevent potential accident.

Reinforcement learning is recapped here to facilitate problem formulation. It is inspired by human's goal-directed learning process and modelled as an agent interacts with the environment $\mathcal{E}$ by balancing between the exploration and exploitation for a maximum return G (the accumulation of reward r_t) in the long run. The agent's decision on the action a_t is based on the current state s_t of the environment by referring to the policy π. It requires that such a sequential decision-making process is in accordance with the Markov decision process (MDP) for mathematical rigour, which backs the convergence of the general iterative optimization process [142]. To evaluate different policies, the state function V and value function Q are defined in the following:

$$Q_\pi(s,a) = \mathbb{E}_\pi[R_t|S_t = s, A_t = s] \tag{6.1}$$

$$V_\pi(s) = \mathbb{E}_{a\sim\pi(s)}[Q_\pi(s,a)] \tag{6.2}$$

$$Q_\pi(s,a) = \sum_{s',a'} p(s'|s,a)\,\pi(a|s)[r + Q_\pi(s',a')] \tag{6.3}$$

One step induction of (6.1) leads to the formula (6.3), which obeys the Bellman equations [142]. Knowing the Q value of each pair (s,a) for each policy π is not the target, the purpose is by policy iteration to find the optimal policy π^* satisfying (6.4) and (6.5):

$$Q_{\pi^*}(s,a) = \max_\pi Q_\pi(s,a) \tag{6.4}$$

$$V_{\pi^*}(s) = \max_\pi V_\pi(s) \tag{6.5}$$

There are several methods in achieving this; however, for brevity it directly goes to the Q-learning which is an off-policy temporal difference algorithm [142]:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) +$$
$$\alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right] \tag{6.6}$$

The intuition here is that the learned Q-function directly approximates the optimal function Q_{π^*}, independent of the policy being followed. Notably, conventional Q-learning is only capable of solving problems with limited state space and action choice. For complicated problems with large state space, one idea is to map state space to action preference by using a function. This is the work systematically investigated by DeepMind [43, 143-145]. Optimization in this way is specifically governed by the following formula (6.7):

$$L_t(\theta_t) = \mathbb{E}_{s,a\sim\rho(\cdot)}\left[\left(y_t - Q(s, a;\, \theta_t)\right)^2\right] \tag{6.7}$$

$$y_t = \mathbb{E}_{s'\sim\mathcal{E}}\left[r + \gamma \max_{a'} Q(s', a';\, \theta_{t-1})\right] \tag{6.8}$$

where Q is the action-value function parameterised by θ. $\rho(\cdot)$ indicates the trajectory distribution, and $\mathcal{E}$ indicates the state distribution in a certain environment. DeepMind contributions in optimizing it are readdressed later.

To adjust the driving safety problem resulting in a solution from reinforcement learning algorithms, a retrospect of supervised learning is presented to highlight the challenges. For supervised learning, the EEG data in the baseline region are extracted and coupled with the measured RT for training and testing. Exemplified by the work in [146, 147], in Fig. 6.1, only EEG signals covered by the blue masks are utilized during the analyzing procedures, because it is ideal to extract brain dynamic there [146]. Most EEG data covered by green masks are discarded due to the lack of label or RT information, which means insufficient data utilization. However, when deploying the trained model for application, the captured EEG data are continuously input for inference about the mind state. There is no differentiation of EEG data segments from the temporal perspective. Such a wider generalization scope can incur poor performance due to the discrepancy in the distributions among EEG data for training and for application.

The paradigm of reinforcement learning requires abstraction and instantiation of the agent, environment, state, action and reward to target a specific problem, which is assumed to comply with MDP. The critic for driving safety can be alternatively attributed to the good prediction of RT, based on which appropriate actions can be taken. The criterion for the quality of predicted RT is its deviation from the corresponding measured RT. When comparing the two, the less the deviation, the better the quality. Therefore, a mechanism is needed to maintain the predicted RT in order to compare with the measured RT when necessary (i.e., where there exists measured RT for the current state). To summarize, the environment can be designed as comprising the

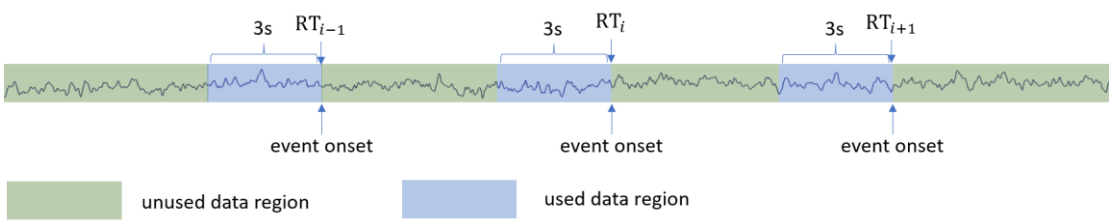


Fig. 6.1 Data utilization paradigm of supervised learning. Only EEG data under the blue mask are used for training and validation.

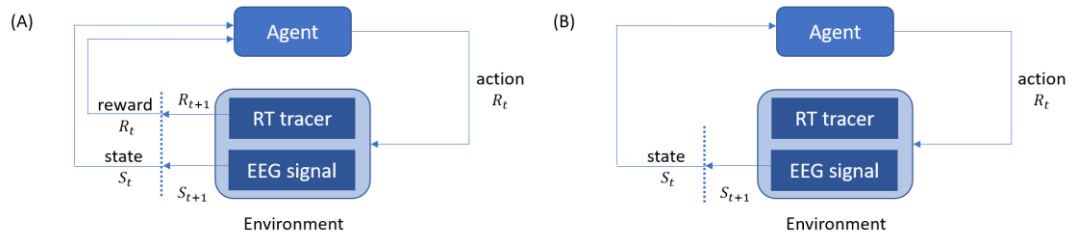


Fig. 6.2 Reinforcement learning paradigm for RT prediction (A) The training stage; (B) The test stage.

captured EEG data and an internal RT tracer. The agent issues actions to manipulate the tracer to estimate RT. This strategy leads to the design of a reinforcement learning architecture in Fig. 6.2.

An instantiation of the reinforcement learning process based on the above depiction is illustrated in Fig. 6.3. The components of reinforcement learning are dissected into a concrete representation in the context of session data captured in one experiment. As shown in Fig. 6.3, the state s_t is one piece of the continuously segmented session data, and the action a_t indicates how to operate the internal tracer. To fabricate a reward at each time step, the following strategy is taken. If the duration of state s_t covers a measured RT, r_t is equal to the negative of the absolute difference between the measured RT (mRT) and the traced RT (tRT); otherwise, the reward r_t is equal to 0. Note that the negative of the absolute value is taken because the optimization targets a maximum return. Because it is focused on Q-learning, the operation to the internal tracer is specially designed to allow only discrete operations, which is intrinsically required by the algorithm itself [142]. As shown in Fig. 6.3, an exemplified action can be chosen to maintain the current traced RT, or to increase/decrease the RT by a certain unit such as 0.5s. Based on the above description, the interaction between the agent and environment is as follows: at time step t, by analyzing the EEG data segment s_t, the agent acts according to a_t to operate tracer with a consequent reward r_t. The traced value gets updated which is transparent to the agent, and environment evolves into the new state s_{t+1}.

It is pointed out that although the actions designed above are self-evident from the illustrative perspective, a mechanism called RT proposition is employed upon implementation. The problem for the action paradigm in Fig. 6.3 is that the final predicted RT by tracing is history-dependent. Although it is a more general case than Markov-dependent, it might not be flexible enough. Roughly, it can be postulated that mind state at $t+1$ only depends on its predecessor at t. Empirical experience can justify this: supposing that someone is sleepy at time step t, he/she can either feel sleepy or experience sudden alertness at $t+1$. Either way, it is not necessary to check the mind state at $t-1$. This argument leads to the formulation of the up-to-date traced RT value as follows:

$$\mathrm{tRT}_t = \beta \cdot \mathrm{tRT}_{t-1} + (1-\beta) \cdot \mathrm{pRT}_t \tag{6.9}$$

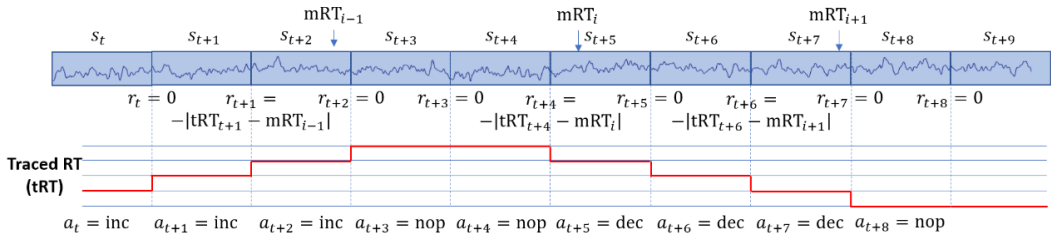


Fig. 6.3 Instantiation of concepts in reinforcement learning in the context of driving safety research.

In (6.9), pRT denotes the proposed RT at the current time step t and can only have discrete values such as 1.2s, 2.5s, etc., due to Q-learning is considered here as aforementioned. β denotes the tradeoff between previous tRT and current pRT and is named transition weight. Because traced RT is termed from the illustrative dynamic decision-making perspective, it is used interchangeably with the term predicted RT which is more commonly used in supervised learning. An extra note is made here that the predicted RT is different from the proposed RT, the latter is only internal to the agent. pRT bid by agent via different actions is used to modulate tRT, which is equivalent to the predicted RT unless otherwise specified.

Notably, compared with the supervised learning, which focuses on trial data, reinforcement learning in this part works directly with session data. It is mentioned that the currently available dataset might pose a challenge for the proposed method, because session data are not abundant due to the original experiment design. However, as long as the practicability can be demonstrated via the current dataset, it still indicates a promising beginning for subsequent research. It is also pointed out that the design is from the prototyping perspective, it might lack the field-engineering rigor. For example, the designed state is the non-overlap EEG segment in 3 seconds. The treatment is just derived from lab convention and might not be the optimal choice in real application.

6.2.2 Network Design

DeepMind has performed revolutionary work in revising conventional reinforcement learning methods, such as Q-Learning. They devised deep Q-learning, or DQN if the underlying DNN for the Q-function estimation needs to be emphasized [43], which was originally thought to be impractical due to several constraints. They are the first to introduce several tricks such as target network, experience replay, etc., which pave the way for successfully applying DQN in different fields. They also enhanced DQN by incorporating more techniques from traditional reinforcement learning and proposing new architectures such as double DQN and duelling DQN [144, 145]. This work adopts the same tricks during the utilization of DQN and makes some modifications to the experience replay to achieve more efficient computation. As shown in Fig. 6.4, a batch buffer is added that refers to the frames in the replay queue to avoid unnecessary copying operations. A sequence number is added which in theory can wrap back to 0 for each frame, to adjust the sample selection.

To design a DQN that is suitable for analyzing EEG data for action or RT proposition, previous work in [56] are referred to guide overall architecture. The principles include less pre-processing of input data to relieve EEG domain knowledge requirement, effective computation to facilitate online deployment, similar input-output

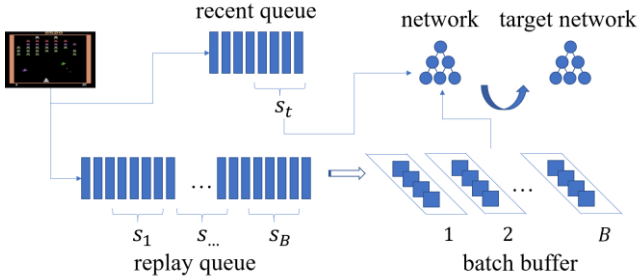


Fig. 6.4 Modification of the experience replay.

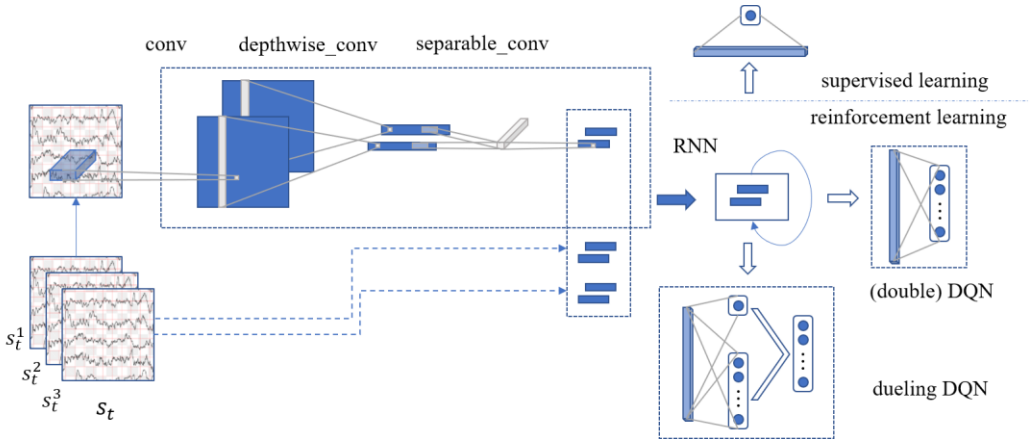


Fig. 6.5 Network architecture. Note that up to RNN, these components can be shared by both the supervised learning and reinforcement learning paradigms. For reinforcement learning, DQN or double DQN, can share the structure up to RNN with dueling DQN.

interface to recycle learning algorithms with DQN and its variants. These considerations lead to the designed Q-network architecture shown in Fig. 6.5.

The proposed network architecture is elaborated on in a certain degree to ease the understanding of some technical details. First, the dataflow up to the RNN part (inclusively) is identical to the recurrent convolutional neural network (RCNN) structure as in [71]. It utilizes a weight-sharing CNN to extract spatial features for subsequent processing by RNN to further explore the temporal information. Accordingly, the input to the network are pre-processed multi-channel EEG data. As mentioned above, the session data are segmented into successive chunks in a certain unit, which is 3 seconds (s) in this paper. For step t, the data s_t which is sliced into three pieces of 1s length, aka s_t^1, s_t^2 and s_t^3 are consecutively fed into the network for feature extraction. These convolutional operations constituting the CNN are arranged in a dashed box to indicate that the weights are shared between processing s_t^1, s_t^2 and s_t^3. The details of the variants of convolutional operation such as depthwise and separable convolution can be found in [88, 148]. After extracting relevant features via CNN, these features are stacked together to feed into the RNN for further processing.

Then, the dataflow sprang from RNN are conducted to multiple network components for different purposes, as depicted in Fig. 6.5. The subnetwork for supervised learning is commonly seen in literatures and hence self-explanatory here [33]. By adding a fully connected layer with one neuron, the network in this circuitry is used as a regressor to predict RT.

For (double) DQN in this work, it adopts the similar treatment in [43], which is essentially by employing DNN as Q-function to boost the performance of Q-learning. In brevity, suppose $s \in \mathcal{S} \subset \mathbb{R}^n$, $a \in \mathcal{A} = \{1, 2, \cdots, |\mathcal{A}|\}$, the constructed network is not in line with the Q-function definition which maps $[\mathbb{R}^n, \mathcal{A}]$ to $\mathbb{R}$, but instead maps $\mathbb{R}^n$ to $\mathbb{R}^{|\mathcal{A}|}$ from the implementing perspective. Consequently, DQN is constructed as a two-layered perceptron, with one layer used to adjust the dimension and a final layer generating Q value for each action. Further computations are based on the output from network, such as employing max operator to select the action to interact with the environment.

For dueling DQN, the insight is that to estimate the value of each action for some states is unnecessary. So, two sub-networks are constructed to separately estimate value function V and advantage function A in accordance with (6.10):

Table 6.1 Network architecture configuration

Model	Layer ID	Layer Type	Filters	Size	Act.	Pad.	Comment
CNN	1	Conv2D	32	(1, 64)	-	Same	
	2	Depthwise-Conv2D	1	(30, 1)	tanh	Valid	Weight clipping
		AvgPool2D	-	(2, 2)		Same	
	3	Separable-Conv2D	32	(1, 16)	tanh	Same	
		AvgPool2D		(2, 2)		Same	
RNN	4	Conv2D	4*3*32	(1, 8)	-	Same	2D LSTM
Supervised		Linear	-	512	id		L_2 Regularization
		Linear		1			
(Double) DQN		Linear	-	512	id		L_2 Regularization
		Linear		#actions			
Dueling DQN	state	Linear	-	512	id		L_2 Regularization
		Linear		1			
	state-action	Linear	-	512	id		L_2 Regularization
		Linear		#actions			

$$Q(s,a,;\theta,\alpha,\beta) = V(s,a;\theta,\alpha) + A(s,a;\theta,\beta) \tag{6.10}$$

The definition of A, the advantage function which relates V and Q is as in (6.11):

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s) \tag{6.11}$$

V, Q and A in (6.10) are parameterized version of (6.1), (6.2) and (6.11), and they share part of the network, as reflected by the parameter θ. In this paper, two streams concerning V and A are both implemented as MLP. As in Fig. 6.5, the branch with one neuron is used to estimate V, and the other branch is for A. The two estimations are added together as Q, which bears later stage computation for the functionality of the whole duelling network. The final network configurations for different models are detailed in Table 6.1.

With the designed Q-network, the procedures for EEG data analysis can be summarised into the following steps:

- Carry out the experiment to obtain session data
- Establish the measured information for reward calculation
- Setup parameters like segment length and input the data into the model for training
- Test the performance of the model for application

6.3 Experiments and Results

EEG data captured during the experiment in section 2.3.1 is used to evaluate the proposed method. As aforementioned, one trait of EEG data that tends to affect the performance is the intra- and cross-subject variance. For experiment, the single-subject and the multi-subject case are separately considered. Because the DQN is firstly initiated for this process oriented BCI paradigm especially from an application perspective, the discrete action space requires limited RT resolution (due to discrete actions) to approach the measured RT, which in theory is a real number. This restriction causes difficulty in competing with the existing results achieved by very deep network structures especially due to quantization error. In addition, the pre-processing of the EEG data is rather lightweight, hence the main consideration here is not to seek the best network structure to compete with the state-of-the-art result. Instead, by comparing with a supervised learning model that shares the network structure as shown in Fig. 6.5, it is

targeted to demonstrate the feasibility and practicability of this new application-oriented computation paradigm for driving performance research via EEG.

EEG signals are in general complicated due to signal distortion, artefact pending, etc. In this work, only limited pre-processing is applied to the data and measured RT. For EEG signals, a band-pass filter with a range 0.5 to 50 Hz is introduced, followed by down-sampling to 128 Hz from the original 500 Hz sampling rate. For RT, the values are clipped into the range 0.5 to 8 second (s), with a moving average within 90 seconds [84]. Finally, the post pre-processed continuous signals are segmented into consecutive sections with a unit of 3 s as shown in Fig. 6.3.

6.3.1 Single-Subject Case

The variance of the statistical distributions for EEG data of a single subject is not as severe as that of multiple subjects in general. In light of this fact, the single subject case is considered first to verify the performance of the proposed model. The subjects chosen are with the same gender and had participated in the experiment at least three times. For some sessions, the experiment was configured to issue warnings upon event onsets, to inspect the influence of explicit arousal on driving performance. These sessions are omitted for consistent experimental scenarios. The subjects that merit the above criteria are S01(4)(1410), S09(3)(704) and S41(3)(1405). The number in the first parentheses indicates number of session data, aka the number of times the subject participated in the experiment. The number in the second parentheses indicates total trials of all sessions. For each subject, the data from the first two sessions are for training and the remaining one is for testing. One session of data for S01 is used for validation to decide total number of iterations which is applied for training on all subjects' data.

The dueling DQN architecture is utilized due to its performance and train the model for 2000 iterations or episodes with similar hyperparameters adopted from [43]. The transition weight β is set to 0.75. To monitor the convergence of the training process of the proposed model, two indicators are employed for inspection: the episode return and the average return. The episode return is the summation of rewards at each step when processing session data. It is a direct measurement of episodic gain but tends to be noisy. By introducing the average return over episodes, the trend for convergence is relatively clearer. However, due to the accumulation effect, there might be a latency of reflection for overfitting. Nevertheless, by combining the two indicators, the training process can be well monitored. The data of two sessions are switched between at random during the training process, and Fig. 6.6 illustrates the corresponding episode return and average return. The trained model is applied to the test session data to predict RT in a consecutive way.

As for the benchmarking method, the same network structure is employed to construct the regressor and train in supervised learning way, according to Fig. 6.5 and

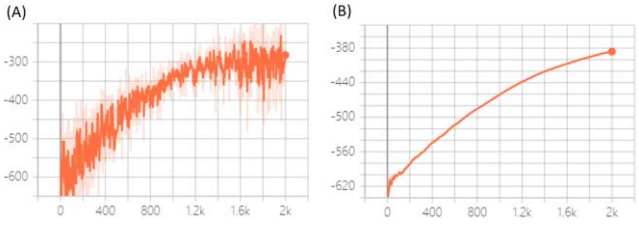


Fig. 6.6 Convergence indicators: (A) the episode return (B) the average return. Convergence indicators: (A) the episode return (B) the average return. The x-axis indicates the sequence of training iteration or episode. The y-axis indicates the return per episode from different perspective.

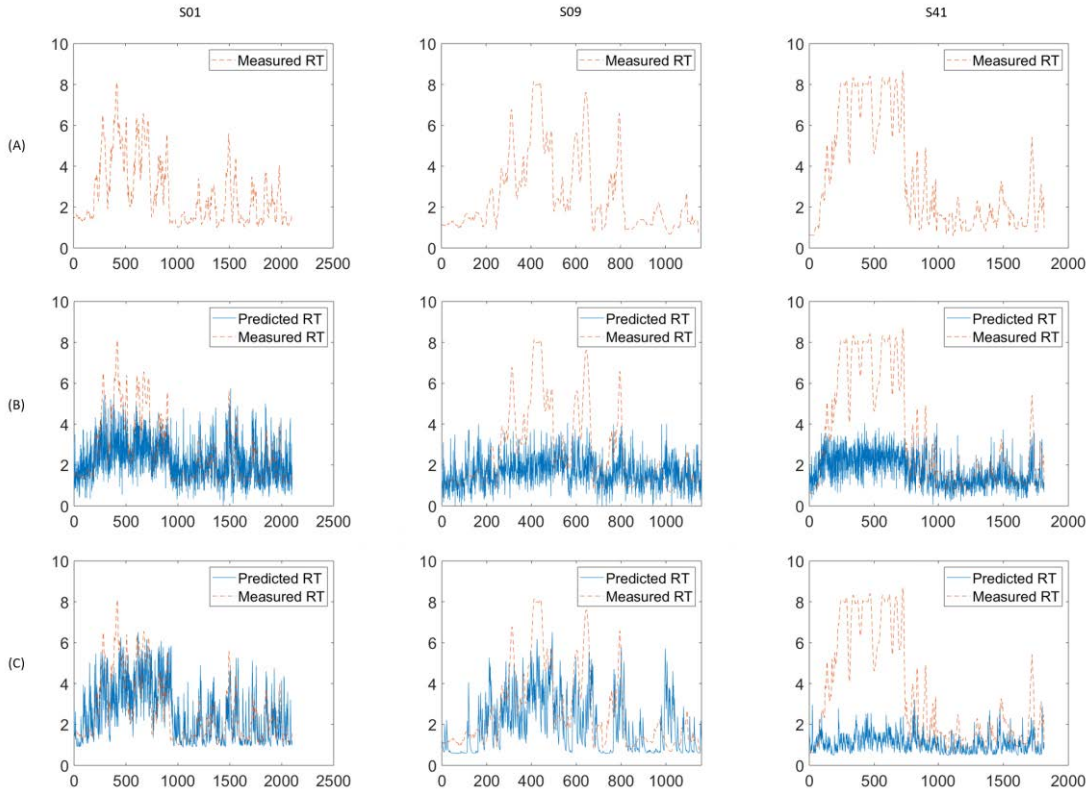


Fig. 6.7 The measured RT vs predicted RT for single-subject test (A) measured RT; (B) supervised learning case; (C) reinforcement learning case. The x-axis indicates the progress of the session (3 seconds per step). The y-axis indicates lengh of RT (unit: seconds).

Table 6.1. The pre-processing of EEG data is the same as the procedures for reinforcement learning. For each trial, 3 seconds duration of EEG signals in the baseline region are extracted as in Fig. 6.1. Coupled with the measured RT, these input/label pairs are used for training and testing respectively. To be consistent with the data division for training and testing as the case of reinforcement learning, trial data of the first two sessions are for training. Trial data of one session of S01 are for validation to decide the total number of iterations during training. The model is trained with learning rate 0.0001 for 600 iterations for each subject. The trained model is tested on trial data to predict RT and also made inference on the test session in a consecutive way as the case of reinforcement learning.

To quantitively assess the performance of reinforcement learning and supervised learning in this case, first, only segments of the test session data where measured RTs exists are considered. The rooted mean square errors (RMSE) between the measured RT and the predicted RT are calculated for the reinforcement learning model and the supervised learning model respectively. Second, to investigate the overall consistency between measured RT and predicted RT, all segments of test data are interpolated based on the measured RT using spline functions and the correlation

Table 6.2 RMSE and correlations of RL and SL for the single-subject case (Unit: seconds)

	Subjects	**S01**	**S09**	**S41**
RMSE	SL	1.26	1.68	**1.87**
	RL	**1.19**	**1.50**	2.29
Correlation	SL	0.50	0.26	0.20
	RL	**0.62**	**0.55**	**0.35**

SL: supervised learning; RL: reinforcement learning

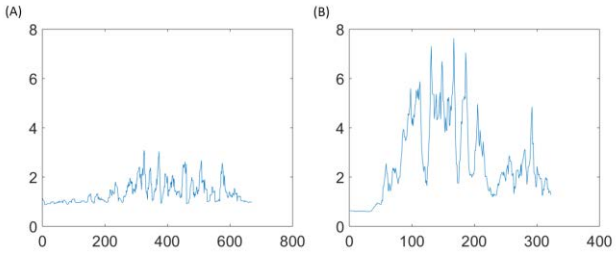


Fig. 6.8 RT distributions of two sessions (A) and (B). The x-axis indicates the sequence of RT in the session. The y-axis indicates lengh of RT (unit: seconds).

with predicted RTs are computed for each model. Fig. 6.7 illustrates the test cases for involved subjects, and Table 6.2 shows the corresponding statistics.

From Table 6.2, it is manifest that with almost the same network architectures but different learning paradigms, the reinforcement learning potentially outperforms supervised learning in most cases. For S01 and S09, reinforcement learning has lower RMSEs but higher correlations. For S41, although the RMSE for reinforcement learning is not as good as supervised learning, it still gets a higher correlation coefficient.

The underperformance of the reinforcement learning model on S41 leads us to further investigate the characteristics of the training set to comprehend this behavior. The RT distributions for the data of two sessions involve in training are in Fig 6.8. It is noticed that the RT distribution of the session in Fig. 6.8(A) is eminently different from that in Fig. 6.8(B). Most RT values in Fig. 6.8(A) are less than 2s. It is well known that one issue for Q-learning is the overestimated to underestimated values, especially when deterministic policy is adopted [144, 149]. During the training process, high chances of encountering small RTs will certainly drive the network to prefer small RT estimations, and this behavior tends to migrate to test stage. This explains the great discrepancy exists between measured RTs and predicted RTs, as observed from the bottom right image of Fig. 6.7. It encourages us to consider more tactics to mitigate the tendency of discrepancy for reinforcement learning as perspective.

6.3.2 Cross-subject Case

For cross-subject performance assessment, subjects that meet the same requirements as for the single-subject case but participated in the experiment one or two times are considered. The subjects that satisfy these criteria are S48(1)(350), S49(2)(705), S50(1)(361), S52(1)(239), and S53(2)(754). The numbers in the parentheses are of the same meanings as in the single-subject case. A leave-out test paradigm is used here. In detail, one session data from all subjects except a specific subject are aggregated for training, and the session data from the given subject for testing. The session data for validation are chosen only from the subjects that had participated in the experiments two times, and one of these two session data are aggregated for validation.

Usually the validation set shares an identical statistical distribution as the training set, although this property cannot be strictly satisfied here. This convention encourages data from S49 and S53 are used for training and validation only. For convenience, data from S48, S50 and S53 are applied for testing. For example, if the one session data of S48 is used for test, then each session from S49, S50, S52 and S53 are collected for training. The left one session from S49 and S53 respectively are collected for validation, typically to decide the number of training iterations.

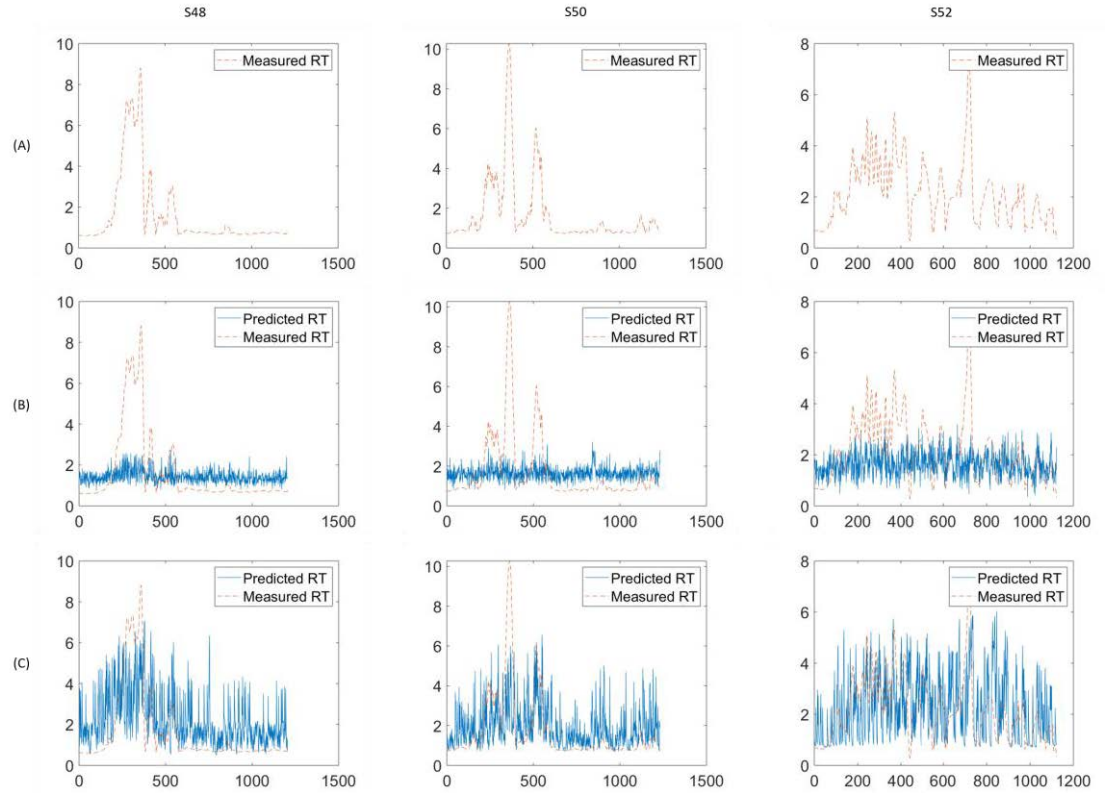


Fig. 6.9 The measured RT vs predicted RT for the cross-subject test: (A) measured RT; (B) supervised learning case; (C) reinforcement learning case. The x-axis indicates the progress of the session (3 seconds per step). The y-axis indicates lengh of RT (unit: seconds).

Furthermore, in addition to the basic pre-processing of EEG data, no normalization or other tricks are adopted to mitigate the impact of cross-subject variance. To reduce the potential risk of overfitting and guarantee good generalization, the double Q-network is adopted to analyze data from multiple subjects, because compared with the dueling Q-network, double Q-network has fewer parameters.

Following the similar procedures and adopting the same hyperparameters as those for the single-subject case (for subject S50, the transition weight is reduced from 0.75 to 0.6), Table 6.3 shows the resulting statistics for both supervised learning and reinforcement learning, and Fig. 6.9 shows the predicted and measured RT of the test data.

From Table 6.3, the RMSE of the prediction of supervised learning is better than the reinforcement learning counterpart. However, as mentioned before, the discrepancy of distribution between training data and test data makes the generalization a challenging task for supervised learning. Reinforcement learning is more robust in this regard as reflected by the correlation coefficients. This means that the model tries to trace the alteration of RT in an overall effective way, which is more significant from an application perspective.

The above results, especially the high correlation coefficients, demonstrate the effectiveness of introducing DQN for EEG data analysis from the session level. If calibration is not a problem, this method can be a suitable candidate for BCI application especially for the single-subject case. In addition, it is postulated that if the experiment is redesigned from the multiple-session perspective to have more session data to train the network, the results can be further enhanced.

Table 6.3 RMSE and correlations of RL and SL for cross-subject case (Unit: second)

	Subjects	S48	S50	S52
RMSE	SL	**1.31**	**1.10**	**1.19**
	RL	1.57	1.16	1.21
Correlation	SL	0.39	0.07	0.16
	RL	**0.48**	**0.45**	**0.33**

SL: supervised learning; RL: reinforcement learning

One trait of the predicted RT of reinforcement learning in Fig. 6.7 and Fig. 6.9 is the high variance of prediction, which is more obvious than the supervised learning counterpart. This is due to a known effect of the deterministic policy gradient [143], which is a limited case of the stochastic policy gradient, providing the variance parameter $\sigma = 0$. However, σ greater than zero is quite common in practice, which leads to high variance when approximating the optimal policy. How to reduce the variance is another direction to improve the performance of the proposed model in future work.

6.4 Discussion and Future Work

First, some background of the research is discussed in section. Currently only Q-learning, specifically DQN is explored to target the problem. Notice in Fig. 6.2(A), the action which is issued to regulate the internal RT tracer has no effect on the mind state. Or alternatively speaking, the transition from state s_t to s_{t+1} is unconditioned on a_t. Because model-based RL tries to predict the next state transition based on the previous state and action, the problem formulation is not feasible for model-based RL. For model-free RL, the solutions include tabular methods represented by Q-learning, approximating methods represented by actor-critic, etc. The latter is not suitable here due to dimension constraint. Consider the navigation problem in [150], based on the raw input image, the state can be abstracted to position and head direction. These latent representations of state can be easily combined with limited action choices in a balanced way and consequently fed into critic network. However, for EEG data, there lacks such a meaningful abstraction. The feature dimension is still quite large compared with number of choices, which is impractical to concatenate together. This constraint brings the problem down to a DQN solution only. To investigate other RL methods is planned as one of the future works.

Next, some interest technical detail concerning the experiment is unveiled and discussed. Although research has clearly revealed the positive correlation between mind state and RT [146, 151], there is no explicit formula to express the relation between the two. Even if this formula exists, due to the current unmeasurable mind state, the correctness of such a formula would be under dispute. However, if some findings in previous work could be consolidated with the research in this paper, the outcomes would be interesting and beneficial.

In [152], the authors indicated that the distinguishable change in mind state is beyond 4 minutes. Referring to the transition weight β in (6.9), to catch the alteration of RT, the initial setting of β is 0.2, to merit the proposed RT based on the current state

Table 6.4 RMSE and correlations for different transition weight values

β	0.2	0.4	0.6	0.75	0.8
RMSE	1.63	1.28	1.20	1.19	1.15
Correlation	0.40	0.61	0.62	0.62	0.60

(EEG data). As revealed in Table 6.4, poor performance seems indicate a false impression that reinforcement learning might not work. However, a review of the related research urges us to try greater β values. The current network structure indicates that $\beta = 0.75$ is a preferred value, which is a good balance between the RMSE and correlation coefficient, even if it is not sure it is the optimal value when the mind state transition is modelled as in (6.9).

Any value of β larger than 0.5 indicates that the network favours the traced RT historically maintained over the newly proposed RT. Although the measured RT tends to be problematic and noisy, it is still a good indicator of the procedural mind state alterations. By aligning RT with EEG data via the model, it consolidates the conclusion that mind state changes in a procedural way [152, 153], which can guide the design of a better apparatus for driving safety. For example, if the current mind state is regarded as alert, a sudden prolonged predicted RT should not incur some warning, because it tends to be a false alarm.

7. Summaries

This research first went through certain neuroscience preliminaries concerning EEG, to emphasize the characteristic of EEG data which bring difficulty to conventional ML algorithms. Then the evolvement of neural networks up to the current deep version was briefly reviewed to unveil its connection with neuroscience and its latest advancement. And based on three stereotypical EEG experiments, which also represent stereotypical BCI paradigms, DNN was investigated from many aspects for EEG data analysis, with set goals encompassed by a research map.

To understand the characteristics of different network structures regarding the trait of EEG data, the research studied certain artificial neural network structures and biological neural network structures by comparison. This inspired the design of more efficient neural architectures, which motivates the construction of a recurrent residual network. The tailored network showed the best result among different models against EEG driving data. Moreover, by seeking to the connection with conventional signal processing, frequency responses of learned filters were well interpreted. They are in accordance with the achievements via statistical analysis from neurophysiological perspective.

To mitigate the intra-subject and cross-subject variance which bring difficulty in generalizing learned models, a subject adaptation network was proposed. It is based on the idea of GAN to seek coherent latent representations among EEG data of different subjects. The quality and coherence of the representation is illustrated by comparing with the designed artificial distribution which is of nice properties, such as clear separations between modalities. This method was used to select more appropriate samples for analyzing the underlying physiological process and illustrating coherent finding. It was also used for classification and came with better result.

To seek the potentiality of utilizing more complicated neural network which incorporates more biological evidence such as memory, memory networks were studied to check the performance in EEG data analysis. In instantiation, DNC was disserted in detail and modified to processing EEG topographical data. A new interpretation of the memory operation led to a stacked version of DNC, which achieved best results among all models on two EEG datasets. As the subsequent work in this regard, SOM was proposed as external memories. It could relieve two challenges which complicate DNC, i.e., the addressing and allocation problems. By incorporating SOM with CNN, it demonstrated improved performances on several datasets.

The last part of this research focused on new computation paradigm, specifically, reinforcement learning, to analyze EEG data of driving. The purpose was to relieve the requirement of high-quality label. Monitoring mind state during driving via EEG also represents an important BCI paradigm. In practice, it means EEG is not correlated to some local event or stimulus, but to the procedural change. Based on the examination, a DQN is specially tailored in tracing the drowsiness level. Results showed the prestige of this method, which presented higher correlation with measured RT on test dataset. This research encouraged more consideration of sophisticated deep reinforcement learning techniques in future research.

To sum it up, by a series of studies and papers, this research addressed certain basic gaps in introducing DNN for EEG data analysis. The conclusions range from network structure selection and architecting to new computation paradigm in processing EEG data, as well as solutions to mitigate high data variance. However, the EEG data from either clinical or practical perspective demands better interpretation to the model behavior and results, which is still of limited achievements. Future work lies in the aspects that to continuously improve or incorporate DNN theories and practice to target EEG data captured in a wider range, as well as explanations of these accomplishments in the context of neuroscience or BCI scenarios.

Reference

[1] "Human Brain." https://en.wikipedia.org/wiki/Human_brain (accessed 05/12, 2018).

[2] L. R. Squire, *Encyclopedia of neuroscience* (Gale virtual reference library). Amsterdam ; Boston: Academic Press, 2009.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015, doi: 10.1038/nature14539.

[4] M. F. Bear, *Neuroscience : exploring the brain*, Fourth edition. ed. Philadelphia: Wolters Kluwer, 2016.

[5] G. Sten *et al.*, "Worldwide initiatives to advance brain research," *Nature Neuroscience,* vol. 19, no. 9, p. 1118, 2016, doi: 10.1038/nn.4371.

[6] E. Commission. "BRAIN RESEARCH." http://ec.europa.eu/research/health/pdf/brain-research_en.pdf (accessed 05/12, 2018).

[7] "BRAIN Initiative." https://www.braininitiative.nih.gov/ (accessed 05/12, 2018).

[8] "Human Brain Project." https://www.humanbrainproject.eu/en/ (accessed 05/12, 2018).

[9] M. Demitri. "Types of Brain Imaging Techniques." https://psychcentral.com/lib/types-of-brain-imaging-techniques/ (accessed 01/06, 2018).

[10] B. Farnsworth. "EEG vs. MRI vs. fMRI – What are the Differences." https://imotions.com/blog/eeg-vs-mri-vs-fmri-differences/ (accessed 01/06, 2018).

[11] E. Allde. "UPDATED – Neuroimaging: EEG, MRI, fMRI, MEG, PET and TMS." http://cognitiveconsonance.info/2014/01/13/updated-neuroimaging-eeg-mri-fmri-meg-pet-and-tms/ (accessed 01/06, 2018).

[12] E. Niedermeyer, *Electroencephalography : basic principles, clinical applications, and related fields*, Fifth edition. ed. Philadelphia: Lippincott Williams & Wilkins, 2005.

[13] S. Sanei, *EEG signal processing* (Electroencephalography signal processing.). Chichester, England ;: John Wiley & Sons, 2007.

[14] E. M. Mizrahi, *Atlas of neonatal electroencephalography*, Fourth edition. ed. New York: Demos Medical Publishing, LLC, 2016.

[15] D. L. Schomer, London. Principles of electroencephalography.

[16] S. N. Abdulkader, A. Atia, and M.-S. M. Mostafa, "Brain computer interfacing: Applications and challenges," *Egyptian Informatics Journal,* vol. 16, no. 2, pp. 213-230, 2015, doi: 10.1016/j.eij.2015.06.002.

[17] S. Saha, K. Mamun, G. Naik, A. Khandoker, S. Darvishi, and M. Baumert, "Progress in Brain Computer Interfaces: Challenges and Trends," *arXiv.org,* 2019.

[18] M. Hajinoroozi, Z. Mao, Y.-P. Lin, and Y. Huang, "Deep Transfer Learning for Cross-subject and Cross-experiment Prediction of Image Rapid Serial Visual Presentation Events from EEG Data," in *Augmented Cognition. Neurocognition and Machine Learning*, Cham, D. D. Schmorrow and C. M. Fidopiastis, Eds., 2017// 2017: Springer International Publishing, pp. 45-55.

[19] Z. Yin and J. Zhang, "Cross-subject recognition of operator functional states via EEG and switching deep belief networks with adaptive weights," *Neurocomputing,* vol. 260, pp. 349-366, 2017/10/18/ 2017, doi: https://doi.org/10.1016/j.neucom.2017.05.002.

[20] "10/20 System Positioning Manual." Trans Cranial Technologies Ldt. https://www.trans-cranial.com/docs/10_20_pos_man_v1_0_pdf.pdf (accessed 01/12, 2017).

[21] A. Savelainen. "An introduction to EEG artifacts." http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.4108/pdf-files/esav10.pdf (accessed 01/07, 2017).

[22] P. G lisse, *Awake and sleep EEG : activation procedures and artifacts*, Second edition. ed. (Atlas of Electroencephalography ; Volume 1). Surrey, England: John Libbey Eurotext, 2016.

[23] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, "Seperability of four-class motor imagery data using independent components analysis," *Journal of Neural Engineering,* vol. 3, no. 3, pp. 208-216, 2006, doi: 10.1088/1741-2560/3/3/003.

[24] M. Salvaris and F. Sepulveda, "Classification effects of real and imaginary movement selective attention tasks on a p300-based braincomputer interface," *Journal of Neural Engineering,* vol. 7, no. 5, p. 056004, 2010, doi: 10.1088/1741-2560/7/5/056004.

[25] R. Parloff. "From 2016: Why Deep Learning Is Suddenly Changing Your Life." Fortune. http://fortune.com/ai-artificial-intelligence-deep-machine-learning/ (accessed 01/05, 2017).

[26] G. Lewis-Kraus. "The Great A.I. Awakening." The New York Times Magzine. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html (accessed 01/05, 2017).

[27] W. Pitts and W. S. McCulloch, "How we know universals the perception of auditory and visual forms," *The bulletin of mathematical biophysics,* vol. 9, no. 3, pp. 127-147, 1947/09/01 1947, doi: 10.1007/BF02478291.

[28] W. S. McCulloch and W. J. T. b. o. m. b. Pitts, "A logical calculus of the ideas immanent in nervous activity," vol. 5, no. 4, pp. 115-133, 1943.

[29] R. Rosenblatt, "Perceptron: a Perceiving and Recognizing Automaton," 1957.

[30] D. O. Hebb and D. Hebb, *The organization of behavior*. Wiley New York, 1949.

[31] S. Hochreiter and J. J. N. c. Schmidhuber, "Long short-term memory," vol. 9, no. 8, pp. 1735-1780, 1997.

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. J. P. o. t. I. Haffner, "Gradient-based learning applied to document recognition," vol. 86, no. 11, pp. 2278-2324, 1998.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[34] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012, doi: 10.1109/MSP.2012.2205597.

[35] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *J. Mach. Learn. Res.,* vol. 12, no. null, pp. 2493–2537, 2011.

[36] "Chemicals; Study Findings on Chemicals Are Outlined in Reports from Merck & Company (Deep neural nets as a method for quantitative structure-activity relationships)," ed. Atlanta: NewsRx, 2016, p. 2422.

[37] T. Ciodaro, D. Deva, J. M. de Seixas, and D. Damazio, "Online particle detection with neural networks based on topological calorimetry information," *Journal of Physics: Conference Series,* vol. 368, no. 1, p. 012030, 2012, doi: 10.1088/1742-6596/368/1/012030.

[38] D. Bychkov *et al.*, "Deep learning based tissue analysis predicts outcome in colorectal cancer," *Sci Rep,* vol. 8, no. 1, pp. 3395-3395, 2018, doi: 10.1038/s41598-018-21758-3.

[39] C. Shallue and A. Vanderburg, "Identifying Exoplanets with Deep Learning: A Five Planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90," *arXiv.org,* vol. 155, no. 2, 2017, doi: 10.3847/1538-3881/aa9e09.

[40] D. Hubel and T. Wiesel, "David Hubel and Torsten Wiesel," *Neuron,* vol. 75, no. 2, pp. 182-184, 2012, doi: 10.1016/j.neuron.2012.07.002.

[41] W. Torsten. "The Neural Basis of Visual Perception." http://centennial.rucares.org/index.php?page=Neural_Basis_Visual_Perception (accessed 01/02, 2018).

[42] I. J. Goodfellow *et al.*, "Generative adversarial nets," presented at the Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, Montreal, Canada, 2014.

[43] M. Volodymyr *et al.*, "Human-level control through deep reinforcement learning," *Nature,* vol. 518, no. 7540, p. 529, 2015, doi: 10.1038/nature14236.

[44] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," presented at the Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France, 2015.

[45] R. M. Stephen *et al.*, "A role for cortical nNOS/NK1 neurons in coupling homeostatic sleep drive to EEG slow wave activity," *Proceedings of the National Academy of Sciences,* vol. 110, no. 50, p. 20272, 2013, doi: 10.1073/pnas.1314762110.

[46] C. Chang, Z. Liu, M. C. Chen, X. Liu, and J. H. Duyn, "EEG correlates of time-varying BOLD functional connectivity," *NeuroImage,* vol. 72, pp. 227-236, 2013, doi: 10.1016/j.neuroimage.2013.01.049.

[47] N. M. Long, J. F. Burke, and M. J. Kahana, "Subsequent memory effect in intracranial and scalp EEG," *NeuroImage,* vol. 84, pp. 488-494, 2014, doi: 10.1016/j.neuroimage.2013.08.052.

[48] N. Pal *et al.*, "EEG-Based Subject- and Session-independent Drowsiness Detection: An Unsupervised Approach," *EURASIP Journal on Advances in Signal Processing,* vol. 2008, no. 1, pp. 1-11, 2008, doi: 10.1155/2008/519480.

[49] A. Telpaz, "Using EEG to predict consumers' future choices," *Journal of marketing research,* 2015.

[50] L.-W. Ko *et al.*, "Sustained Attention in Real Classroom Settings: An EEG Study.(Report)," *Frontiers in Human Neuroscience,* vol. 11, 2017, doi: 10.3389/fnhum.2017.00388.

[51] *Niedermeyer's electroencephalography : basic principles, clinical applications, and related fields*, Sixth edition. ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health, 2011.

[52] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE,* vol. 89, no. 7, pp. 1123-1134, 2001, doi: 10.1109/5.939829.

[53] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed. ed. (Computer science and scientific computing.). Boston: Academic Press, 1990.

[54] A. Kai Keng, C. Zhang Yang, Z. Haihong, and G. Cuntai, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," vol. 10, ed: IEEE, 2008, pp. 2390-2397.

[55] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Muller, "Optimizing Spatial filters for Robust EEG Single-Trial Analysis," *IEEE Signal Processing Magazine,* vol. 25, no. 1, pp. 41-56, 2008, doi: 10.1109/MSP.2008.4408441.

[56] 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," (in eng), *J Neural Eng,* vol. 15, no. 5, p. 056013, Oct 2018, doi: 10.1088/1741-2552/aace8c.

[57] O. Dressler, G. Schneider, G. Stockmanns, and E. F. Kochs, "Awareness and the EEG power spectrum: analysis of frequencies," *BJA: British Journal of Anaesthesia,* vol. 93, no. 6, pp. 806-809, 2004, doi: 10.1093/bja/aeh270.

[58] B. J. Roach and D. H. Mathalon, "Event-Related EEG Time-Frequency Analysis: An Overview of Measures and An Analysis of Early Gamma Band Phase Locking in Schizophrenia," *Schizophrenia Bulletin,* vol. 34, no. 5, pp. 907-926, 2008, doi: 10.1093/schbul/sbn093.

[59] D. A. Bridwell *et al.*, "Moving Beyond ERP Components: A Selective Review of Approaches to Integrate EEG and Behavior.(electroencephalograhy)(Report)(Brief article)," *Frontiers in Human Neuroscience,* vol. 12, 2018, doi: 10.3389/fnhum.2018.00106.

[60] T. Yi-Li, K. Chii-Shyang, and M. Liou, "Event-related spectral perturbations and phase synchronization of EEG during natural auditory memory retrieval task," ed: IEEE, 2014, pp. 1-4.

[61] A. H. Mooij, B. Frauscher, M. Amiri, W. M. Otte, and J. Gotman, "Differentiating epileptic from non-epileptic high frequency intracerebral EEG signals with measures of wavelet entropy," *Clinical Neurophysiology,* vol. 127, no. 12, pp. 3529-3536, 2016, doi: 10.1016/j.clinph.2016.09.011.

[62] J. N. Acharya, A. Hani, P. Thirumala, and T. N. Tsuchida. "Guideline 3: A Proposal for Standard Montages to Be Used in Clinical EEG." American Clinical Neurophysiology Society. https://www.acns.org/UserFiles/file/EEGGuideline3Montage.pdf (accessed 01/12, 2017).

[63] J. Britz, D. Van De Ville, and C. M. Michel, "BOLD correlates of EEG topography reveal rapid resting-state network dynamics," *NeuroImage,* vol. 52, no. 4, pp. 1162-1170, 2010, doi: 10.1016/j.neuroimage.2010.02.052.

[64] A. Bersagliere, R. Pascual-Marqui, L. Tarokh, and P. Achermann, "Mapping Slow Waves by EEG Topography and Source Localization: Effects of Sleep Deprivation," *Brain Topography,* vol. 31, no. 2, pp. 257-269, 2018, doi: 10.1007/s10548-017-0595-6.

[65] B. Saletu, P. Anderer, and G. M. Saletu-Zyhlarz, "EEG Topography and Tomography (LORETA) in Diagnosis and Pharmacotherapy of Depression," *Clinical EEG and Neuroscience,* vol. 41, no. 4, pp. 203-210, 2010, doi: 10.1177/155005941004100407.

[66] M. Cosottini *et al.*, "EEG topography-specific BOLD changes: a continuous EEG-fMRI study in a patient with focal epilepsy," *Magnetic Resonance Imaging,* vol. 28, no. 3, pp. 388-393, 2010, doi: 10.1016/j.mri.2009.11.011.

[67] G. Ungureanu, C. Bigan, R. Strungaru, and V. Lazarescu, "Independent Component Analysis Applied in Biomedical Signal Processing," vol. 4, 11/30 2003.

[68] P. Margaux, M. Emmanuel, D. Sébastien, B. Olivier, and M. Jérémie, "Objective and Subjective Evaluation of Online Error Correction during P300-Based Spelling," *Advances in Human-Computer Interaction,* vol. 2012, 2012, doi: 10.1155/2012/578295.

[69] K. Robbins, K.-M. Su, and W. D. Hairston, "An 18-subject EEG data collection using a visual-oddball task, designed for benchmarking algorithms and headset performance comparisons," *Data Brief*, vol. 16, pp. 227-230doi: 10.1016/j.dib.2017.11.032.

[70] C.-H. Chuang, L.-W. Ko, T.-P. Jung, and C.-T. Lin, "Kinesthesia in a sustained-attention driving task," *NeuroImage,* vol. 91, pp. 187-202, 2014, doi: 10.1016/j.neuroimage.2014.01.015.

[71] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," *arXiv.org,* 2016.

[72] O. A. Rosso *et al.*, "Wavelet entropy: a new tool for analysis of short duration brain electrical signals," *Journal of Neuroscience Methods,* vol. 105, no. 1, pp. 65-75, 2001, doi: 10.1016/S0165-0270(00)00356-3.

[73] M. Fiecas and H. Ombao, "Time Series Modeling of Neuroscience Data," *Journal of the American Statistical Association,* vol. 109, no. 508, p. 1719, 2014.

[74] S. Kumar, A. Sharma, K. Mamun, and T. Tsunoda, "A Deep Learning Approach for Motor Imagery EEG Signal Classification," ed: IEEE, 2016, pp. 34-39.

[75] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals," *Computers in Biology and Medicine,* vol. 100, pp. 270-278, 2018, doi: 10.1016/j.compbiomed.2017.09.017.

[76] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping,* vol. 38, no. 11, pp. 5391-5420, 2017, doi: 10.1002/hbm.23730.

[77] M. van Putten, S. Olbrich, and M. Arns, "Predicting sex from brain rhythms with deep learning," *Sci Rep,* vol. 8, no. 1, pp. 3069-3069, 2018, doi: 10.1038/s41598-018-21495-7.

[78] A. Graves *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature,* vol. 538, no. 7626, pp. 471-476, 2016/10/01 2016, doi: 10.1038/nature20101.

[79] C. Shorten and T. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data,* vol. 6, no. 1, pp. 1-48, 2019, doi: 10.1186/s40537-019-0197-0.

[80] W. M. Kouw, "An introduction to domain adaptation and transfer learning," *ArXiv,* vol. abs/1812.11806, 2018.

[81] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," ed: IEEE, 2017, pp. 1-6.

[82] G. Zhang, K. K. W. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Analysis and Prevention,* vol. 87, p. 34, 2016, doi: 10.1016/j.aap.2015.10.033.

[83] J. Odom *et al.*, "Visual evoked potentials standard (2004)," *Documenta Ophthalmologica,* vol. 108, no. 2, pp. 115-123, 2004, doi: 10.1023/B:DOOP.0000036790.67234.22.

[84] W. Dongrui, V. J. Lawhern, and B. J. Lance, "Reducing Offline BCI Calibration Effort Using Weighted Adaptation Regularization with Source Domain Selection," ed: IEEE, 2015, pp. 3209-3216.

[85] B. Alan, "Working memory: looking back and looking forward," *Nature Reviews Neuroscience,* vol. 4, no. 10, p. 829, 2003, doi: 10.1038/nrn1201.

[86] P. Bashivan, G. M. Bidelman, and M. Yeasin, "Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity," *European Journal of Neuroscience,* vol. 40, no. 12, pp. 3774-3784, 2014, doi: 10.1111/ejn.12749.

[87] A. V. Oppenheim, *Discrete-time signal processing*, 3rd ed. ed. (Prentice Hall signal processing series.). Upper Saddle River [N.J: Pearson, 2010.

[88] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *arXiv.org,* 2017.

[89] K. Cho, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv.org,* 2014.

[90] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv.org,* 2015.

[91] G. Huang, Z. Liu, and K. Weinberger, "Densely Connected Convolutional Networks," *arXiv.org,* 2018.

[92] Y. Lu *et al.*, "Revealing Detail along the Visual Hierarchy: Neural Clustering Preserves Acuity from V1 to V4," *Neuron,* vol. 98, no. 2, pp. 417-428.e3, 2018, doi: 10.1016/j.neuron.2018.03.009.

[93] *Principles of neural science*, 5th ed. ed. New York: McGraw-Hill, 2013.

[94] A. Fonseca, S. Kerick, J.-T. King, C.-T. Lin, and T.-P. Jung, "Brain Network Changes in Fatigued Drivers: A Longitudinal Study in a Real-World Environment Based on the Effective Connectivity Analysis and Actigraphy Data," *Frontiers in Human Neuroscience,* vol. 12, 2018, doi: 10.3389/fnhum.2018.00418.

[95] F. Ruyi, A. Kai Keng, Q. Chai, G. Cuntai, and W. Aung Aung Phyo, "An analysis on driver drowsiness based on reaction time and EEG band power," vol. 2015, ed: IEEE, 2015, pp. 7982-7985.

[96] G. Pfurtscheller and F. H. Lopes Da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology,* vol. 110, no. 11, pp. 1842-1857, 1999, doi: 10.1016/S1388-2457(99)00141-8.

[97] M. Gola, M. Magnuski, I. Szumska, and A. Wróbel, "EEG beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects," *International Journal of Psychophysiology,* vol. 89, no. 3, pp. 334-341, 2013, doi: 10.1016/j.ijpsycho.2013.05.007.

[98] D. M. Roberts, J. R. Fedota, G. A. Buzzell, R. Parasuraman, and C. G. McDonald, "Prestimulus Oscillations in the Alpha Band of the EEG Are Modulated by the Difficulty of Feature Discrimination and Predict Activation of a Sensory Discrimination Process," vol. 26, no. 8, pp. 1615-1628, 2014, doi: 10.1162/jocn_a_00569.

[99] T. A. Rihs, C. M. Michel, and G. Thut, "Mechanisms of selective inhibition in visual spatial attention are indexed by α-band EEG synchronization," *European Journal of Neuroscience,* vol. 25, no. 2, pp. 603-610, 2007, doi: 10.1111/j.1460-9568.2007.05278.x.

[100] J. J. Jiang, "A Literature Survey on Domain Adaptation of Statistical Classifiers," 2007.

[101] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics,* vol. 60, no. 4, pp. 699-746, 2008, doi: 10.1007/s10463-008-0197-x.

[102] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," vol. 69, B. Zadrozny, Ed., ed, 2004.

[103] K. Zhang, B. SchoÏkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," *30th International Conference on Machine Learning, ICML 2013,* pp. 1856-1864, 01/01 2013.

[104] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Transferring Visual Category Models to New Domains," EECS Department, University of California, Berkeley, UCB/EECS-2010-54, May 7 2010. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-54.html

[105] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative Domain Adaptation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 22-29 Oct. 2017 2017, pp. 2784-2792, doi: 10.1109/ICCV.2017.301.

[106] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial Discriminative Domain Adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 2962-2971, doi: 10.1109/CVPR.2017.316.

[107] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 22-29 Oct. 2017 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[108] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv.org,* 2017.

[109] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," *arXiv.org,* 2018.

[110] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter Bank Common Spatial Pattern Algorithm on BCI Competition IV Datasets 2a and 2b," *Frontiers in neuroscience,* vol. 6, pp. 39-39, 2012, doi: 10.3389/fnins.2012.00039.

[111] I. Kant, *The Critique of Pure Reason*. Auckland: The Floating Press, 2009.

[112] Eric r. Kandel, Y. Dudai, and Mark r. Mayford, "The Molecular and Systems Biology of Memory," *Cell,* vol. 157, no. 1, pp. 163-186, 2014, doi: 10.1016/j.cell.2014.03.001.

[113] L. Squire and J. Wixted, "The Cognitive Neuroscience of Human Memory Since H.M," *Annual Review of Neuroscience,* vol. 34, p. 259, 2011.

[114] D. Graupe and H. Kordylewski, "A large scale memory (LAMSTAR) neural network for medical diagnosis," vol. 3, ed: IEEE, 1997, pp. 1332-1335 vol.3.

[115] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing Machines," *arXiv.org,* 2014.

[116] J. Weston, S. Chopra, and A. Bordes, "Memory Networks," *arXiv.org,* 2015.

[117] I. Goodfellow, *Deep learning* (Adaptive computation and machine learning series.). Cambridge, Massachusetts: MIT Press, 2016.

[118] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 26-31 May 2013 2013, pp. 6645-6649, doi: 10.1109/ICASSP.2013.6638947.

[119] G. Karol, I. Danihelka, A. Graves, and D. Wierstra, "DRAW: A Recurrent Neural Network For Image Generation," *arXiv.org,* 2015.

[120] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised Learning for Physical Interaction through Video Prediction," *arXiv.org,* 2016.

[121] M. Prasad, Y. Huang, Y.-K. Wang, and C.-T. Lin, *Brain Dynamic States Analysis based 3D Convolutional Neural Network*. 2017.

[122] V. Bekhtereva, C. Sander, N. Forschack, S. Olbrich, U. Hegerl, and M. M. Müller, "Effects of EEG-vigilance regulation patterns on early perceptual processes in human visual cortex," *Clinical Neurophysiology,* vol. 125, no. 1, pp. 98-107, 2014, doi: 10.1016/j.clinph.2013.06.019.

[123] J. R. Foucher, H. Otzenberger, and D. Gounot, "Where arousal meets attention: a simultaneous fMRI and EEG recording study," *Neuroimage,* vol. 22, no. 2, pp. 688-697, 2004, doi: 10.1016/j.neuroimage.2004.01.048.

[124] S. Olbrich *et al.*, "EEG-vigilance and BOLD effect during simultaneous EEG/fMRI measurement," *NeuroImage,* vol. 45, no. 2, pp. 319-332, 2009, doi: 10.1016/j.neuroimage.2008.11.014.

[125] J.-H. Kim, D.-W. Kim, and C.-H. Im, "Brain Areas Responsible for Vigilance: An EEG Source Imaging Study," *Brain Topography,* vol. 30, no. 3, pp. 343-351, 2017, doi: 10.1007/s10548-016-0540-0.

[126] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, Cham, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014// 2014: Springer International Publishing, pp. 818-833.

[127] A. P. Yonelinas, M. Aly, W. C. Wang, and J. D. Koen, "Recollection and familiarity: Examining controversial assumptions and new directions," vol. 20, J. L. Voss and K. A. Paller, Eds., ed. Hoboken: Wiley Subscription Services, Inc., A Wiley Company, 2010, pp. 1178-1194.

[128] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 79, no. 8, p. 2554, 1982, doi: 10.1073/pnas.79.8.2554.

[129] T. J. B. c. Kohonen, "Self-organized formation of topologically correct feature maps," vol. 43, no. 1, pp. 59-69, 1982.

[130] A. Turing, "The chemical basis of morphogenesis," *Bulletin of Mathematical Biology,* vol. 52, no. 1-2, pp. 153-197, 1990, doi: 10.1007/BF02459572.

[131] C. Malsburg, "Self-organization of orientation sensitive cells in the striate cortex," *Kybernetik,* vol. 14, no. 2, pp. 85-100, 1973, doi: 10.1007/BF00288907.

[132] T. Kohonen, E. Oja, O. Simula, A. Visa, and J. Kangas, "Engineering applications of the self-organizing map," *Proceedings of the IEEE,* vol. 84, no. 10, pp. 1358-1384, 1996, doi: 10.1109/5.537105.

[133] H. Nimmagadda, S. Sukhavasi, and S. Babu, "Self Organising Maps: An Interesting Tool for Exploratory Data Analysis," *Research Journal of Engineering and Technology,* vol. 3, no. 4, pp. VI-326, 2012.

[134] S. Clark, S. A. Sisson, and A. Sharma, "Nonlinear manifold representation in natural systems: The SOMersault," *Environmental Modelling & Software,* vol. 89, p. 61, 2017.

[135] "Molecular mechanisms of learning and memory," *Journal of Neurochemistry,* vol. 88, pp. 17-17, 2004, doi: 10.1046/j..2003.0c4-1.x.

[136] D. L. Schacter and A. D. Wagner, "Learning and memory," in *Principles of neural science (5th Ed.)*, E. R. Kandel, J. R. Schwartz, and T. M. Jessell Eds. New York: McGraw-Hill.

[137] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The PREP pipeline: standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics,* vol. 9, no. JUNE, 2015, doi: 10.3389/fninf.2015.00016.

[138] I. Winkler, S. Haufe, and M. Tangermann, "Automatic Classification of Artifactual ICA-Components for Artifact Removal in EEG Signals," *Behavioral and Brain Functions,* vol. 7, no. 1, p. 30, 2011/08/02 2011, doi: 10.1186/1744-9081-7-30.

[139] D. Powers and Ailab, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *J. Mach. Learn. Technol,* vol. 2, pp. 2229-3981, 01/01 2011, doi: 10.9735/2229-3981.

[140] C.-T. Lin *et al.*, "Minority Oversampling in Kernel Adaptive Subspaces for Class Imbalanced Datasets," *IEEE Transactions on Knowledge and Data Engineering,* vol. 30, no. 5, pp. 950-962, 2018, doi: 10.1109/TKDE.2017.2779849.

[141] "World report on road traffic injury prevention," World Health Organization (WHO). [Online]. Available: http://www.who.int/violence_injury_prevention/publications/road_traffic/world_report/chapter3.pdf?ua=1

[142] R. S. Sutton, *Reinforcement learning : an introduction*, Second edition. ed. (Adaptive computation and machine learning series.). Cambridge, Massachusetts: The MIT Press, 2018.

[143] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms," in *ICML*, Beijing, China, 2014-06-21 2014. [Online]. Available: https://hal.inria.fr/hal-00938992. [Online]. Available: https://hal.inria.fr/hal-00938992

[144] H. v. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-Learning," presented at the Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, 2016.

[145] Z. Wang, T. Schaul, M. Hessel, and M. Lanctot, "Dueling Network Architectures for Deep Reinforcement Learning," *arXiv.org,* 2016.

[146] L. Yu-Ting, L. Yang-Yin, W. Shang-Lin, C. Chun-Hsiang, and L. Chin-Teng, "Brain Dynamics in Predicting Driving Fatigue Using a Recurrent Self-Evolving Fuzzy Neural Network," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 27, no. 2, pp. 347-360, 2016, doi: 10.1109/TNNLS.2015.2496330.

[147] Y. Ming, Y.-K. Wang, M. Prasad, D. Wu, and C.-T. Lin, "Sustained Attention Driving Task Analysis based on Recurrent Residual Neural Network using EEG Data," vol. 2018-, ed: IEEE, 2018, pp. 1-6.

[148] J. Guo, Y. Li, W. Lin, Y. Chen, and J. Li, "Network Decoupling: From Regular to Depthwise Separable Convolutions," *arXiv.org,* 2018.

[149] C. J. C. H. Watkins, "Learning from Delayed Rewards," University of Cambridge, 1989.

[150] A. Banino *et al.*, "Vector-based navigation using grid-like representations in artificial agents," *Nature,* 2018.

[151] C.-H. Chuang, Z. Cao, J.-T. King, B.-S. Wu, Y.-K. Wang, and C.-T. Lin, "Brain Electrodynamic and Hemodynamic Signatures Against Fatigue During Driving.(Brief article)," *Frontiers in Neuroscience,* vol. 12, 2018, doi: 10.3389/fnins.2018.00181.

[152] S. Makeig and M. Inlow, "Lapses in alertness: coherence of fluctuations in performance and EEG spectrum," *Electroencephalography and clinical neurophysiology,* vol. 86, no. 1, pp. 23-35, 1993.

[153] J. Tzyy-Ping, S. Makeig, M. Stensmo, and T. J. Sejnowski, "Estimating alertness from the EEG power spectrum," *IEEE Transactions on Biomedical Engineering,* vol. 44, no. 1, pp. 60-69, 1997, doi: 10.1109/10.553713.