



EEG data analysis with stacked differentiable neural computers

Yurui Ming¹ · Danilo Pelusi² · Chieh-Ning Fang¹ · Mukesh Prasad¹ · Yu-Kai Wang¹ · Dongrui Wu³ · Chin-Teng Lin^{1,4}

Received: 11 July 2018 / Accepted: 9 November 2018 / Published online: 20 November 2018
© Springer-Verlag London Ltd., part of Springer Nature 2018

Abstract

Differentiable neural computer (DNC) has demonstrated remarkable capabilities in solving complex problems. In this paper, we propose to stack an enhanced version of differentiable neural computer together to extend its learning capabilities. Firstly, we give an intuitive interpretation of DNC to explain the architectural essence and demonstrate the stacking feasibility by contrasting it with the conventional recurrent neural network. Secondly, the architecture of stacked DNCs is proposed and modified for electroencephalogram (EEG) data analysis. We substitute the original Long Short-Term Memory network controller by a recurrent convolutional network controller and adjust the memory accessing structures for processing EEG topographic data. Thirdly, the practicability of our proposed model is verified by an open-sourced EEG dataset with the highest average accuracy achieved; then after fine-tuning the parameters, we show the minimal mean error obtained on a proprietary EEG dataset. Finally, by analyzing the behavioral characteristics of the trained stacked DNCs model, we highlight the suitability and potential of utilizing stacked DNCs in EEG signal processing.

Keywords Deep Learning (DL) · Differentiable neural computer (DNC) · Electroencephalogram (EEG) · Stacked DNCs

1 Introduction

The motivation for the invention of differentiable neural computer (DNC) [1] is natural from the neurobiological perspective. It is exemplified by the fact that human beings' high cognitive level 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 believed that the new memory-augmented architecture which has already demonstrated its remarkable capability in solving complex tasks would lead to new horizons [2]. Hence, it is worthwhile to investigate its performance in variant research fields like electroencephalogram (EEG) data analysis.

✉ Chin-Teng Lin
Chin-Teng.Lin@uts.edu.au

Yurui Ming
Yurui.Ming@student.uts.edu.au

Danilo Pelusi
dpelusi@unite.it

Chieh-Ning Fang
Chieh-Ning.Fang@uts.edu.au

Mukesh Prasad
Mukesh.Prasad@uts.edu.au

Yu-Kai Wang
Yukai.Wang@uts.edu.au

Dongrui Wu
drwu@hust.edu.cn

¹ Centre for Artificial Intelligence, University of Technology Sydney, Sydney, NSW, Australia

² Faculty of Communication Sciences, University of Teramo, Teramo, Italy

³ School of Automation, Huazhong University of Science and Technology, Wuhan, China

⁴ Faculty of Engineering and Information Technology, School of Software, University of Technology Sydney, Sydney, Australia

However, EEG data analysis is often quite challenging because of factors such as high artifact pending, cross-subject variance, low signal-to-noise ratio [3–5]. Due to the neurophysiological traits or physical constraints, these adversities are inevitable and impact the selection of the appropriate models in achieving highly accurate results. Hence, it is wise to have a good understanding of DNC to properly exert its potential from an applicational perspective. When tracing the evolution of DNC, from its ancestor the Neural Turing machine (NTM) [6] to its current state, it can be well understood that the interaction with memory is the most critical and most complex part of the whole system. It is best to begin with the approach of memory interaction in order to have an essential interpretation of DNC.

In this paper, we investigate the operation and memory mechanism of DNC and based on these studies the stacked DNCs architecture is proposed. Our contributions lie in several folds as numerated: (1) We give an intuitive interpretation about the way of memory augmentation in DNC, which simplifies the description and inspires the extension. (2) By analogy to multilayer Long Short-Term Memory (M-LSTM), we propose to stack multiple DNCs together to promote the learning capabilities. (3) We enhance the internal structure and operation of DNC in accordance with the paradigm of the EEG data analysis. (4) We properly illustrate the operation of trained stacked DNCs to highlight its suitability to analyze EEG data.

The rest of the paper is organized as follows. First based on the interpretation of memory mechanism and by comparing with M-LSTM, we stack multiple DNCs together to form a new structure. Then we introduce the paradigm and specialties of EEG data analysis and detail several modifications to the original structure following the instantiation of the stacked DNCs architecture. In the experiment section, we demonstrate the elegance when applying stacked DNCs on two EEG datasets by comparing it with other methods. Finally, a discussion is given about the trained stacked DNCs' behavior which is probed by a white noise mono-topography image to inspire a better understanding of the proposed network and its effectiveness for EEG data analysis.

2 Interpretation

Recurrent neural network (RNN), especially the cell type of LSTM, is currently considered the most successful model in utilizing memory mechanism [7–9]. Unlike plain multilayer 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 time axis are illustrated as in Fig. 1a.

Generally during the recurrent procedures, hidden state at time $t - 1$ is concatenated to the input at next time step t for further processing. Usually, there is an assumption of existence in temporal correlation between successive inputs, which is also the motivation to apply RNN. However, it might not be always the case, like in the instance of the copy task example [6]. There the authors believe 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. 1b, the direct link between adjacent time steps will unavoidably incur temporal couplings. The advantage is it can capture the correlation if it does exist for the problem being modeled. The drawback is that such a tight coupling can be futile in some situations. For instance, if the coupling to the current state is some hidden state of several time steps latency, it is probably faded away when passing through these links up until its current state.

By the introduction of LSTM, we know such a restriction is to some extent relieved. 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 ask is 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 given as in Fig. 1c. Compared with RNN, a matrix or cell block exists acting as an external memory (indicated by the disks in the figure) which characterizes the specialty 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 . Such a modification incurs the complexity from spatial perspective; however, the temporal dependency is dramatically reduced due to memory access methods. The links and interactions between unfolded hidden layers of DNC's controller are shown as in Fig. 1d, 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, as for DNC, an intuitive interpretation could be stated that it is targeted to solve the temporal coupling in traditional RNN. Such an explanation is preferable than some exceedingly general ones. One way 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 weakly dependence

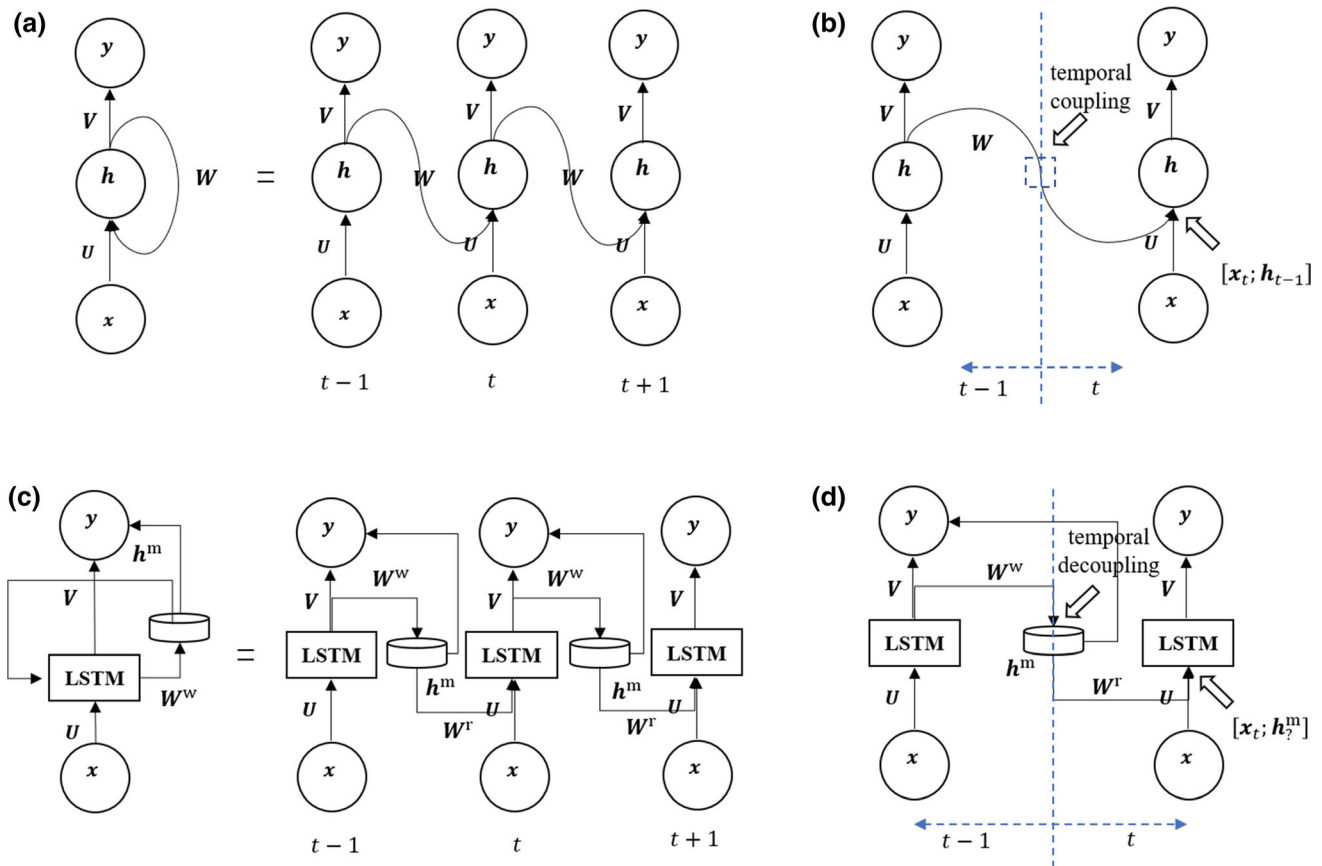


Fig. 1 **a** RNN and its equivalent unfolded version along the time axis, **b** Demonstration of tight temporal coupling in RNN, **c** equivalence of unfolded DNC across time axis, **d** demonstration of temporal decoupling by augmented memory. Note x , y and h are for input, output and hidden state, respectively. We use capital letter U , V and

W to denote weight matrices. For denotation h_t^m in **d**, the question mark indicates due to asynchronous operation, it is usually not known the hidden state at which step involved in the current operation (latent time step)

between consecutive inputs, can be considered to utilize DNC for modeling.

It is now quite natural to stack multiple DNCs together, since such an arrangement just resembles the multilayer RNN or LSTM network which are ubiquitous in deep learning applications [7–9]. A direct comparison is given below as in Fig. 2. The reason that in Fig. 2 for stacked DNCs the recurrent path and external memory are both in dotted lines is to emphasize they are actually internal to DNC not external. The current drawing is for analogous purpose.

There are several advantages to consider stacking DNC together. The most obvious one is by doing so that the scale of hyperparameters gets increased. Though currently the correspondence between complexity of the problem with the magnitude of parameters of the model is still unavailable, it is believed one factor contributing to the success of DL is the sufficient large parameters to form a potentially large representation space. The second benefit is via stacking we can construct a more powerful heterogeneous system. As mentioned in the original paper, the controller

of DNC is free to choose. For multilayer LSTM, although the configuration of each layer can be different from each other, practice from implementation tends to put

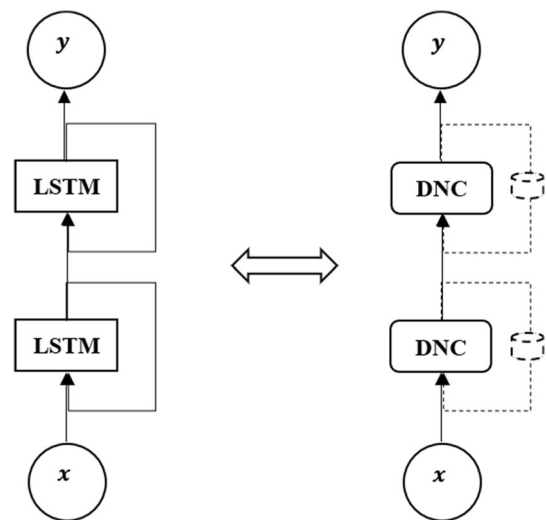


Fig. 2 Comparison of multilayered LSTM and DNC

constraints on such 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. 2 is a more flexible heterogeneous system. Actually, in the application part, the two DNCs we stacked together are slightly different from one another, but by cooperation harmonically, they can form a more capable system than functioning alone.

3 Architecture

Brain research is among the most significant researches currently [10], and one of the promising applications is mind state monitoring via electroencephalogram (EEG) signals captured by noninvasive sensors. The most critical problem here is EEG signals are usually immensely complicated, causing the design of feature-extraction procedure to be very difficult. The general neural networks, particularly the deep version, has proven their ability for auto feature extraction [11]. It is understood such capability is underpinned by the weights being learnt. Sometimes not only the quality but also the quantity of weights decides the performance of the model, and that is why sometimes these parameters are called hyperparameters. For EEG signals, it is quite natural to try memory network since the extra memory mechanism (read-heads and write-heads) to some extent adds to the number of weights to learn and motivates us to stack multiple DNCs together.

3.1 Computation graph

For EEG signal processing, usually a series of EEG segments are related to a certain mind state or cognitive phenomenon and treated by associating pairs. By terming them input and label respectively to free from the neurological context, the problem formed in EEG domain is delivered to the machine learning domain to solve. To regard it as time series data and model it via RNN, it is obvious of a many-to-one mapping paradigm. The original DNC structure in [1] is of the many-to-many type with a dependency graph drawn as in Fig. 3a. So, the first step for EEG application is to deduce the formula for the corresponding many-to-one mapping.

We take the same symbols as in [1], and there for many-to-many mapping, the computation graph is governed by the following formulas:

$$\chi_t = [x_t; r_{t-1}^1; \dots; r_{t-1}^R] \quad (1)$$

$$v_t = W_y[h_t^1; \dots; h_t^L] \quad (2)$$

$$\xi_t = W_\xi[h_t^1; \dots; h_t^L] \quad (3)$$

$$y_t = v_t + W_r[r_t^1; \dots; r_t^R] \quad (4)$$

We leave other formulas like the ones governing the update of the controller network to the next section; for instance, how to compute h_t^i from χ_t and h_t^{i-1} , etc. We only draw a simplified version as in Fig. 3a correspondingly for demonstration.

Since for many-to-one mapping the learning signal only exists for the last time step, so v_t is only necessary for the last time step T . This reasoning leads to the following formulas (5, 6) and the corresponding computation graph as in Fig. 3b.

$$\xi_t = W_\xi[h_t^1; \dots; h_t^L] \quad (5)$$

$$y_T = W_y[h_T^1; \dots; h_T^L] + W_r[r_T^1; \dots; r_T^R] \quad (6)$$

3.2 Controller network

The next concern of designing a stacked DNCs for EEG data analysis lies in the variant input formats that the research community adopts. As in Fig. 4, EEG data can be treated in time domain, frequency domain, or spatial domain depending on the underlying models or algorithms. The most natural way to present EEG data is taking the waveform format from time domain. This is the way most EEG acquisition devices use for storing raw data. However, only with trained eyes and in some special cases, otherwise it is hard to glean information or draw useful conclusion directly from the waveform data [12]. Nevertheless, there is work in directly processing waveform data at the advantage of avoiding intensive preprocessing or frequency domain transform [13].

Frequency domain is conventionally where the EEG analysis is performed [14, 15]. EEG signal oscillation along the time axis is generally the reflection of the underlying brain activities, but noises caused by impedance varies and other factors are unavoidable. By scarifying the time information, some features like frequency distribution and power spectrum are more eminent than the wave bursts or oscillation in time domain. Since transforming from time domain to frequency domain requires EEG data lasting for some duration being processed, it means some unusual variance is discarded at the same time. Based on these global and stable features, more sophisticated biomarkers can be achieved [16].

To study brain function and corresponding connections with different vortex lobes, especially brain activities under some specially designed or induced cognitive process, a convenient way is to examine EEG power alterations of frequency components over spatial domain, aka working

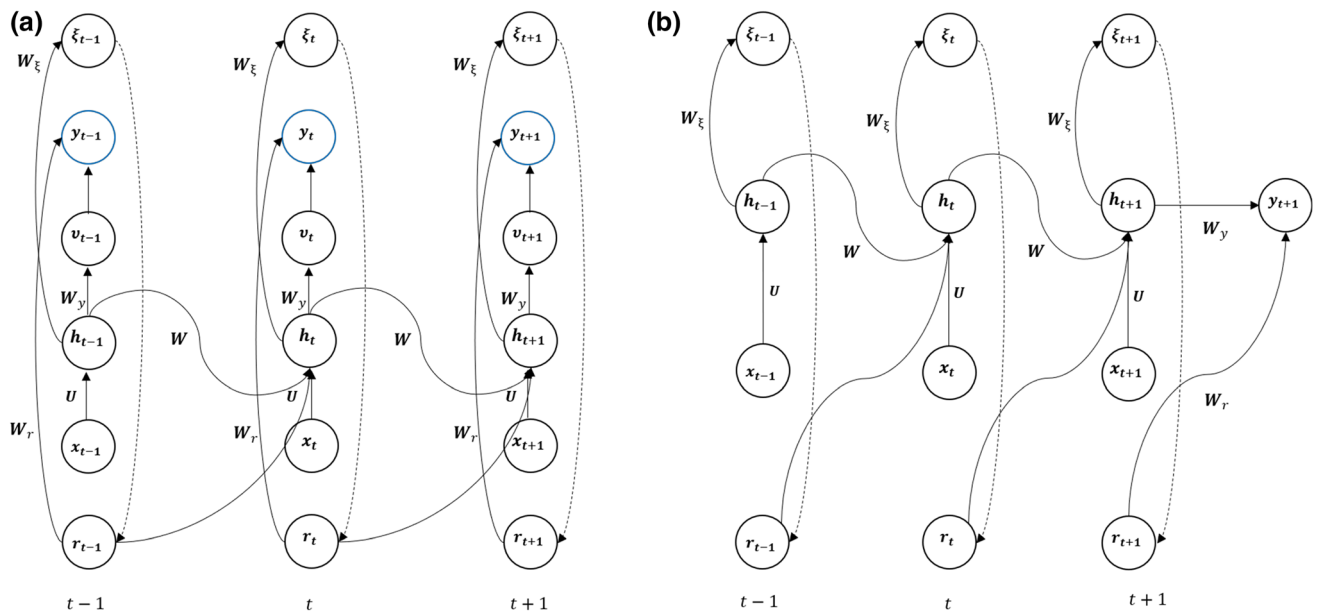


Fig. 3 **a** Computation graph dependency for many-to-many paradigm of DNC, **b** computation graph dependency for many-to-one paradigm of DNC

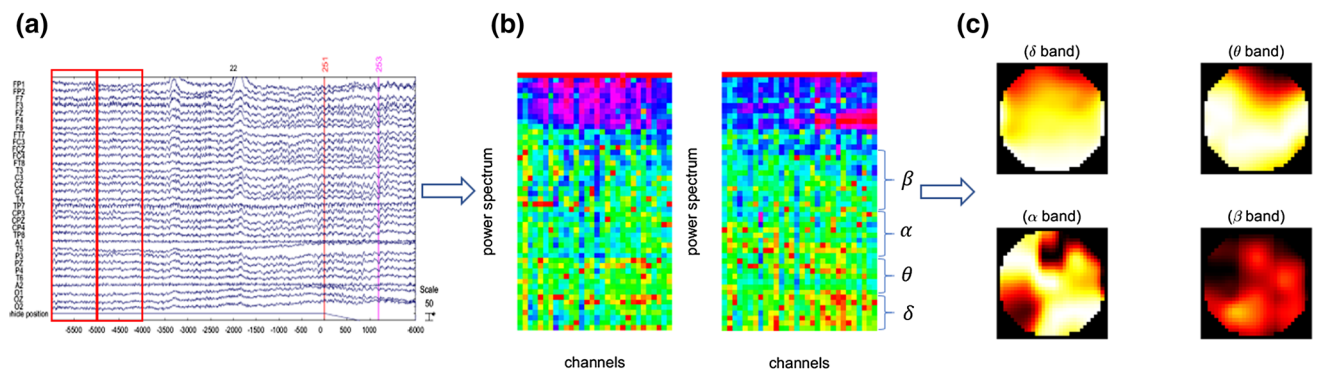


Fig. 4 EEG data format **a** waveform data in time domain, **b** power spectrum in frequency domain (transformation of segmented waveform data), **c** EEG topographies in spatial domain (mapped from power values of same sub-bands of multiple channels). The transformation from **a** to **b** is via Fourier transform for each segment. The

transformation from **b** to **c** is via functionality provided by EEGLab (<https://sccn.ucsd.edu/eeqlab/index.php>). It consists of extrapolation and interpolation based on the available power spectrum to cover the whole scalp meanwhile involving projection from 3D to 2D

with EEG topography [17–19]. It is well known that the success of the deep neural network to some extent attributes to the utilization of convolutional operation iteratively to exploit the local structures of input data. Following this criterion, it is obvious the topographical format for EEG data representation is the most suitable way to work with when considering deep learning. So, in the following we stick to EEG topography to further tailor the network structure.

The relation between different representations or formats of EEG data is also shown in Fig. 4. Note that the transformations are irreversible. In practice the research community puts emphasis mainly on specific frequency bands, aka theta, alpha, beta, respectively, for unveiling or

discussing brain activities and functionalities. When compared with image recognition, the topographical maps of each band added together are identical to color images. Because of this, it is quite direct to apply available tactics in deep learning for processing EEG topographies.

However, in [1], the controller is constructed from a traditional LSTM network which may fail to merit such a spatial correlation between channel values. Inspired by the work in [20], the structure of a recurrent convolutional network is put into consideration here. To effectively blend convolutional operation with recurrent operation and also from 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:

$$\chi_t = \text{concat}(I_t, S_{t-1}, M_{t-1}) \quad (7)$$

$$z_t = \sigma(\text{conv}(\chi_t, W_z)) \quad (8)$$

$$r_t = \sigma(\text{conv}(\chi_t, W_r) + 1.0) \quad (9)$$

$$c_t = \tanh(\text{conv}(\chi_t, W_i)) \quad (10)$$

$$h_t = h_{t-1} * r_t + c_t * z_t \quad (11)$$

In (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 last time step. For addressing similarity, it is preferable to compare between normalized vectors of which no component supremely dominates, so all the activation functions chosen in (8–10) are either sigmoid or hyper-tangent functions. The design of the controller network complying with the above formulas is demonstrated as in Fig. 5a. Overall the controller network can be regarded as a convolutional gated recurrent cell network (Conv.GRU).

3.3 External memory

The last consideration of our proposed architecture lies in the memory layout. The external memory in [1] 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. Since we have to work with topographies which results in all intermediate processing representations essentially being 2D, so 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. 5b. It is pointed out here it is not the optimal way for

exploring the potential of stacked DNCs model, but we will consider a more appropriate version in future work.

3.4 Network architecture

The overall architecture of the network is shown in Fig. 6. In this paper, only two DNCs are stacked together, but there is no constraint on how many DNCs are involved. A maximal pooling layer is also introduced between consecutive DNCs to reduce the dimension, which is a stereotypical treatment in deep neural networks. Since DNC is a general architecture, for instance, the size of external memory is still problem-dependent, so the details of configuration are delayed to the experiment section. However, it is always the architecture as in Fig. 6 we take to analyze different EEG datasets.

4 Experiments

Two EEG datasets are made use of to justify the feasibility and elegance of the proposed stacked DNCs. One is an open-sourced EEG dataset indexing the load of mind [19]; another is our proprietary EEG dataset captured from a sustained-attention driving task [21].

4.1 EEG mind load dataset

To validate the feasibility of our proposed model, the open-sourced data in [19] will be utilized. It is EEG data captured during an experiment which is to measure mind load or cognitive capacity via working memory as in Fig. 7. The test is to have test subjects decide whether a randomly showing letter is belonging to the previously appearing randomly generated letter set or not. For fixed memorizing

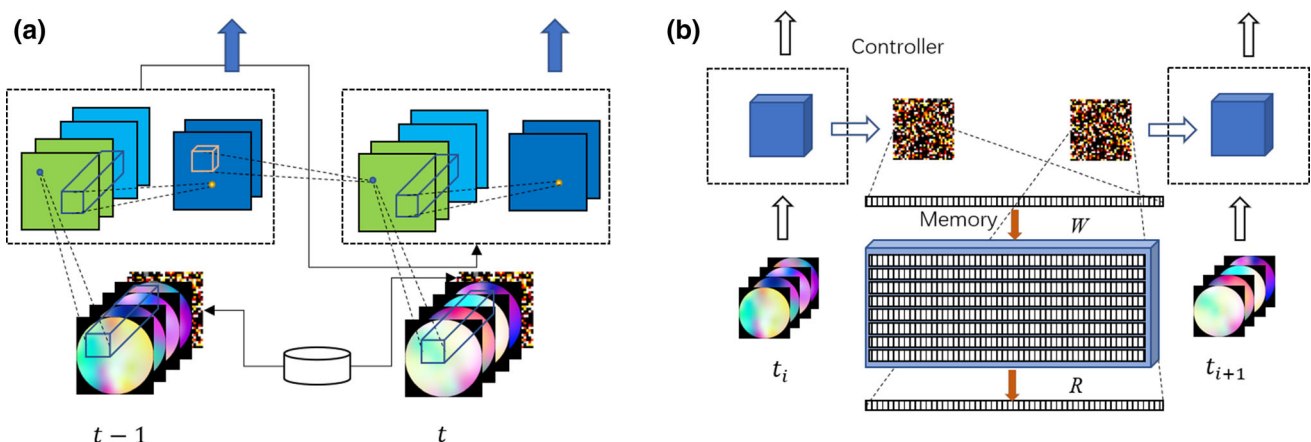


Fig. 5 **a** Controller network constituted by recurrent convolutional network; Interactions with memory is substantially simplified just for indication, **b** specially tailored DNC: a recurrent convolutional

structure substituted the original LSTM; feature maps are flattened and concatenated into vectors to be stored into the external memory

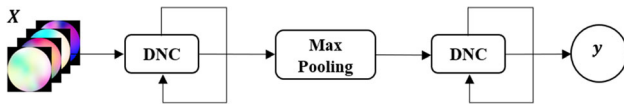


Fig. 6 Overall structure of stacked DNCs; Note the maximum pooling layer between consecutive DNCs

period and retaining period, the mind load is regulated by the size of the letter set, or how many letters are 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. EEG is recorded along the process simultaneously and continuously. Detail of the experiment itself and data capturing specification can be referred to [19].

As in Fig. 7, the duration from display of the randomly generated test letter set to test subjects' respond is called a trial. The EEG data lasting from $t = 1$ to $t = 3500$ ms are taken as input data, which is segmented into 7 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 theta, alpha and beta is put into consideration here. We take the same approach as in [19] by mapping the spectrum components into topographies. The transformation procedure is given in Fig. 4. The size of test set (aka how many letters) is treated as mind load label.

To evaluate the performance by comparing different models, no discrepant treatment is taken compared with the original paper. There were 13 subjects who participated in the experiment. During data preparation, one subject is excluded for each fold. The trials of all left subjects are combined together and shuffled. Then the number of trials which is equal to the excluded subject in the current fold is picked out for cross-validation, and all the remains are used for training. This process is repeated for all the test subjects.

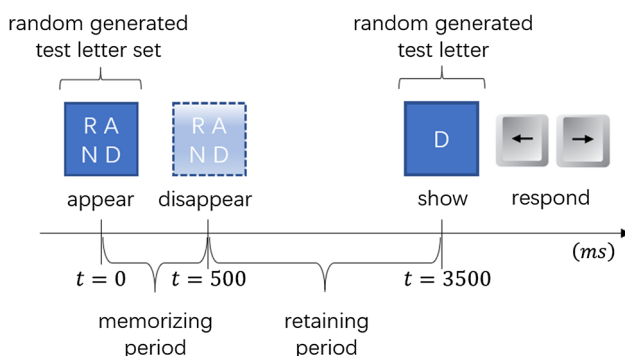


Fig. 7 Mind load via working memory experiment paradigm

We trained the model with the configuration as in Table 1. The batch number is set to 10 percent of training samples which are about 2250. The training process was noticed being convergent after 5 epochs, but the whole duration lasts for 40 epochs. We also take the early stop strategy to prevent overfitting as in the original paper. The statistics for testing and comparison are given in Table 2.

In Table 2, the first three methods are all from the original paper [19]. 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 [19]. The results, however, are still subject dependent.

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

4.2 EEG sustained-attention dataset

Although the mind load dataset is made open-sourced, it is just in the intermediate processed form of the original EEG signal. Namely, they are frequency components of three bands after Fourier transformation to the original pre-processed waveform data. Sometimes, iterations between trying different frequency band composition and test results comparison need to be performed. Therefore, it is more suitable to use an EEG dataset from the very original form to thoroughly evaluate a proposed model. As a result, we

Table 1 Configurations of stacked DNCs for mind load EEG data analysis

Controller		Memory	
<i>DNC1</i>			
Structure	RNN	Slot size	16
Cell type	Conv. GRU	Word size	1024
Filter size	5×5	Read head	4
Feature map	12	Write head	1
<i>DNC2</i>			
Structure	RNN	Slot size	16
Cell type	Conv. GRU	Word size	256
Filter size	5×5	Read head	4
Feature map	12	Write head	1

Table 2 Mind state classification accuracies for each fold and means^a

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

The means (values in bold) are calculated across all subjects' test accuracies

^aPart of the results is directly from [19]

will use a sustained-attention EEG dataset captured during an experiment conducted by us to test the proposed model's performance.

The scenario of the experiment is shown in Fig. 8. The aim of the experiment is to study the relation between fatigue and driving performance, based on the postulate that low vigilance leads to significant latency to various occurrences during driving. During the experiment, recruited test subjects with driver license participated in a simulated driving conducted in a laboratory environment. Subjects would operate a converted car and react to the driving scenario showing before them in giant chained screens. Usually, the car would be cruising along one lane of the highway, then a deliberate perturbation to the car was introduced (deviation onset) and the consequence of car deviation was also synchronized to the display on the screens. The test subjects were instructed to adjust the steering wheel (response onset) to have the car back to the original cruising lane. It is manifest that short reaction time corresponds to alert mind state, while long reaction time corresponds to the mind state of fatigue. The related terminologies are also indicated as in Fig. 8, and a detailed description of the experiment is also contained in [21, 22].

For data preprocessing, the procedures resemble the treatment as in [22]. However, in [22], it works with graphs as shown in Fig. 4b, after manually adjusting the channel sequence for x -axis. As mentioned above, in order to utilize neural networks to automatically explore the spatial relationship between channel locations, it is preferred to work

with topographies as in Fig. 4c. Since the mapping from Fig. 4b and c requires interpolation and extrapolation of original channel values, in this paper all the channels are retained to ensure the interpolating or extrapolating accuracies. It is still common to select some channels according to experience and knowledge to analyze as in [22]. The less picky on EEG data means more information can be kept during transformation. The topographical data preparation procedures are identical to the first experiment above.

For simplicity, the dataset is depicted abstractly in a machine learning context. The preprocessed multichannel waveform EEG data captured during the baseline period (600 ms immediate before deviation onset) after transformation is denoted by X_i . The corresponding reaction time (RT) used to measure the vigilance of test subjects is denoted by y_i . Each trial corresponds to a sample pair (X_i, y_i) . We treat the problem as a regression to predict RT y_i given baseline data X_i . EEG data of five subjects were made use of from balanced-sample perspective.

We take the leave-one-subject-out method as in the first experiment. For a given test subject, the trials from other subjects are used to train the model, and trials for the specified subject are used for testing. With the configuration as in Table 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 five times for each subject and the averaged root-mean-square errors (RMSEs) of the predicted RTs are treated as the final performance indicators.

For comparison, all the models in [19] have been re-implemented as benchmark methods. The statistics are shown in the following Table 4. From the table, it is manifest that the performance achieved by the proposed stacked DNCs is the best among them all. We also compare the correlation coefficients between various models; the values that are overall comparable, which demonstrates the practicability of the model proposed by us. It is believed 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.

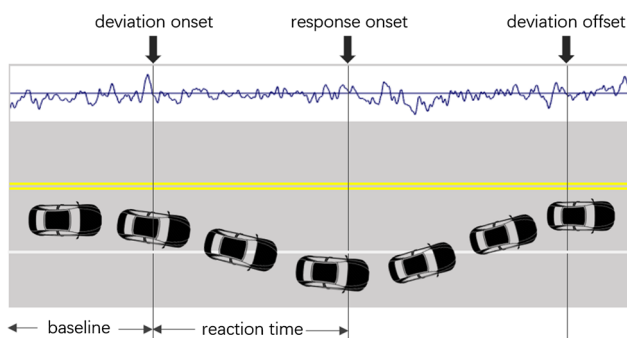
**Fig. 8** Paradigm of sustain-attention driving test

Table 3 Configurations of stacked DNCs for driving EEG data analysis

Controller		Memory	
<i>DNC1</i>			
Structure	RNN	Slot size	32
Cell type	Conv. GRU	Word size	256
Filter size	5×5	Read head	2
Feature map	12	Write head	1
<i>DNC2</i>			
Structure	RNN	Slot size	32
Cell type	Conv. GRU	Word size	256
Filter size	5×5	Read head	2
Feature map	12	Write head	1

5 Discussion

Neural networks are criticized as black boxes in spite of their excellent performance. It is helpful to look into the behavior of the network to have a better understanding of the underlying working principle or to inspire the designing of new architectures. In the following, the second EEG dataset analyzed above, aka the dataset captured during the sustained-attention task, will be focused to carry on the discussion.

The experiment is based on the assumption that the attention of test subjects could not be maintained on the same level during the whole procedure. To measure the attention or fatigue during driving (especially to increase the correlation between attention and reaction time as well as to exclude other impacting factors), participants only need to operate the steering wheel in reaction to the lane-perturbation event and are free from accelerator and brakes pedals controlling. The setup of the experiment is as in Fig. 9.

From the design principles and test setup, it is obvious visuomotor sensory and mind vigilance or attention are key factors of the physiological or cognitive process of the

subjects during the experiment. For visuomotor sensory, due to the structure of the brain [23], it is well known that occipital part will be the most active area [24, 25]. For vigilance, it is believed the frontal cortex is correlated with attention variance [26, 27]. Therefore, it is inclined to expect such priori could be reflected in the trained model.

To inspect the behavior of the stacked DNC after training with our experimental data, we restore the trained parameters and input as a series of identical white noise topography image then observe the manipulated feature maps to seek whether there exist such correlations. The reason for choosing white noise image is we want to know whether for averaged distribution power values, could they be enhanced or saturated by the network especially for some local regions. The feature maps are drawn from the first DNC since it is believed convolutional layers in lower hierarchy of the neural network tends to extract concrete features [11], which is a little easier to make the comparison. Four feature maps selected from a total 12 are drawn in Fig. 10 for demonstration and comparison.

It is interesting to note that feature maps 1 and 2 mainly focus on the occipital area, probably due to signals of visuomotor processes and how they should be specifically captured by the network model [24, 25] for better performance. For feature map 3 and 4, they are focusing on the contralateral frontal part. Since the recruited test subjects are all right-handed and sustained-attention is highly involved in the whole process, it makes sense and has some coincidence with the conclusion as in [26, 27]. So generally speaking, all these parts that got modulated are in accordance to our expectation.

Based on the above observation, we conclude 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, so it tends to be a hard problem to manually extract features from EEG data to help interpret the brain's functionalities for a specific experimental setup. However, such a local concentration mechanism of our model can

Table 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 ^a
	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 ^a
	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 ^a
	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 ^a
	Corr. Coef.	0.339	0.413	0.472	0.389	0.138	0.350

^aMain performance indicators for different models

Fig. 9 Setup of the experiment: **a** converted car from a real one, **b** maneuver platform on which the car is installed, **c** Multiscreen LEDs mimic highway scenario

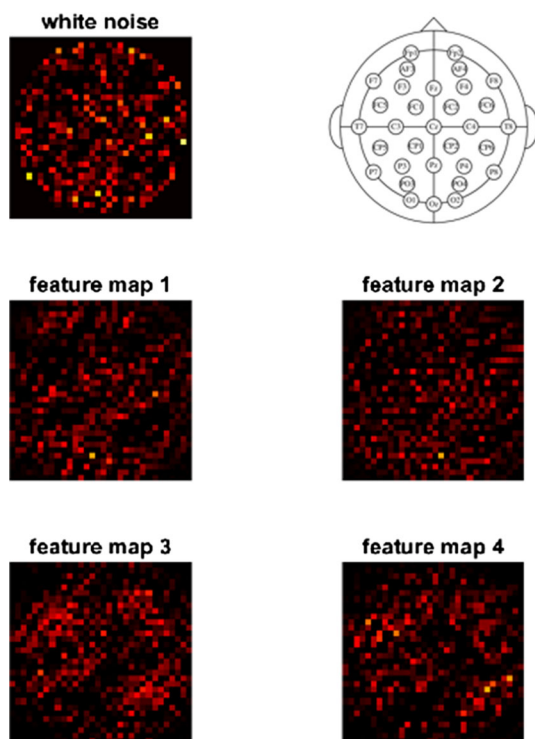
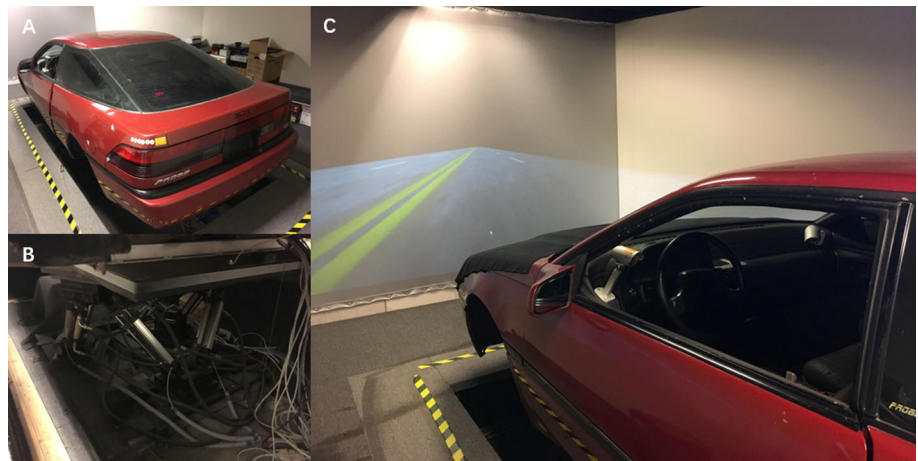


Fig. 10 White noise input and corresponding feature maps. The positional relation in each figure is indicated as in the atlas figure in the top right corner

effectively select the targeted regions in favor of optimal result, which in turn unveils which part of brain is more involved in a specific task than others. Put it together, all these characteristics of DNC especially the stacked one reveal the feasibility and promising potentiality to be utilized for EEG data analysis.

Another aspect worthwhile to mention is the way in which memory gets used. As aforementioned, our modification to the memory-related structure is plausible but may not be the optimal way. One difficulty lies in the implementation of the DNC itself. The original implementation is by

analogous to the modern computer architecture. Although with succeeded applications to complicated problems, following the architecture of modern computers especially the memory part requires two challenges to be solved: 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 of these unavoidably add to the complexity of the whole architecture and make the adjustment or modification to the original memory part extremely difficult. We consider this in future work to address the above problems in some ingenious and direct way in order to drastically reduce the complexity.

6 Conclusion

In this paper, we first compared DNC with the conventional RNN to highlight its asynchronized memory operation characteristics. We then proposed to stack an enhanced version of DNC together based on the intuitive interpretation to expand its representation and learning capabilities. Following that, we gave a brief illustration of EEG data in different formats and highlighted our modification to the internal structure and operation for analyzing EEG topographical data. By applying instantiations of stacked DNCs against EEG datasets captured from various experiments, we showed the best mean results achieved by our model. We gave a short discussion to emphasize the characteristics of stacked DNCs' operation and the potential of utilizing our proposed architecture on EEG data analysis as well as future work to make the interaction with memory part more optimal.

Acknowledgements This work was supported in part by the Australian Research Council (ARC) under discovery Grant DP180100670 and DP180100656. Research was also sponsored in part by the Army Research Laboratory and was accomplished under Cooperative

Agreement Number W911NF-10-2-0022 and W911NF-10-D-0002/TO 0023. The views and the conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Graves A, Wayne G et al (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538:471–476
- Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge
- Lin C-T, Wu R-C, Jung T-P, Liang S-F, Huang T-Y (2005) Estimating driving performance based on EEG spectrum analysis. *EURASIP J Appl Signal Process* 19:3165–3174
- Sanei S (2007) EEG signal processing. Wiley, Hoboken
- Sanei S (2013) Adaptive processing of brain signals. Wiley, Hoboken
- Graves A, Wayne D, Danihelka I (2014) Neural Turing machines. [arXiv:1410.5401](https://arxiv.org/abs/1410.5401)
- Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: IEEE international conference on acoustics, speech and signal processing, pp 6645–6649
- Gregor K, Danihelka I, Alex G, Rezende DJ, Wierstra D (2015) DRAW: a recurrent neural network for image generation. [arXiv:1502.04623v2](https://arxiv.org/abs/1502.04623v2)
- Cho K, van Merriënboer B et al. (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. [arXiv:1406.1078](https://arxiv.org/abs/1406.1078)
- <https://www.braininitiative.nih.gov/>. Accessed 1 May 2018
- LeCun Y, Bengio Y, Hinton G (2015) Deep learning *Science* 321:436–444
- Mizrahi EM (2014) Atlas of neonatal electroencephalography. Lippincott Williams & Wilkins, Philadelphia
- Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ (2017) EEGNet: a compact convolutional network for EEG-based brain–computer interfaces. [arXiv:1611.08024v2](https://arxiv.org/abs/1611.08024v2)
- Konareva I (2005) Modifications of the EEG frequency pattern in humans related to a single neurofeedback session. *Neurophysiology* 37:388–395
- Roach BJ, Mathalon DH (2008) Event-related EEG time–frequency analysis: an overview of measures and an analysis of early gamma band phase locking in schizophrenia. *Schizophr Bull* 34:907–926
- Barry RJ, De Blasio FM (2018) EEG frequency PCA in EEG-ERP dynamics. *Psychophysiology*. <https://doi.org/10.1111/psyp.13042>
- Britz J, Van De Ville D, Michel CM (2010) BOLD correlates of EEG topography reveal rapid resting-state network dynamics. *NeuroImage* 52:1162–1170
- Bersagliere A, Pascual-Marqui RD, Tarokh L, Achermann P (2018) Mapping slow waves by EEG topography and source localization: effects of sleep deprivation. *Brain Topol* 31:257–269
- Bashivan P, Rish I, Yeasin M, Codella N (2016) Learning representations from EEG with deep recurrent-convolutional neural networks. In: International conference on learning representation (ICLR). [arXiv:1511.06448v3](https://arxiv.org/abs/1511.06448v3)
- Finn C, Goodfellow I, Levine S (2016) Unsupervised learning for physical interaction through video prediction. In: Neural information processing systems (NIPS). [arXiv:1605.07157v4](https://arxiv.org/abs/1605.07157v4)
- Chuang C-H, Ko L-W, Jung T-P, Lin C-T (2014) Kinesthesia in a sustained-attention driving task. *NeuroImage* 91:187–202
- Hung Y-C, Wang Y-K, Prasad M, Lin C-T (2017) Brain dynamic states analysis based on 3D convolutional neural network. In: IEEE international conference on systems, man and cybernetics (SMC). <https://doi.org/10.1109/smc.2017.8122606>
- Bear MF, Connors BW, Paradiso MA (2016) Neuroscience: exploring the brain, 4th edn. Wolters Kluwer, Alphen aan den Rijn
- Bekhtereva V, Sander C, Forschack N, Olbrich S, Hegerl U, Müller MM (2014) Effects of EEG-vigilance regulation patterns on early perceptual processes in human visual cortex. *Clin Neurophysiol* 125:98–107
- Foucher JR, Otzenberger H, Gounot D (2004) Where arousal meets attention: a simultaneous fMRI and EEG recording study. *Neuroimage* 22:688–697
- Olbrich S, Mulert C et al (2009) EEG-vigilance and BOLD effect during simultaneous EEG/fMRI measurement. *NeuroImage* 45:319–332
- Kim J-H, Kim D-W, Im C-H (2017) Brain areas responsible for vigilance: an EEG source imaging study. *Brain Topogr* 30:343–351