Applied Soft Computing Journal 84 (2019) 105689

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Subject adaptation network for EEG data analysis

Yurui Ming [a], Weiping Ding [b], Danilo Pelusi [c], Dongrui Wu [d], Yu-Kai Wang [a], Mukesh Prasad [a], Chin-Teng Lin [a,*]

[a] Centre for Artificial Intelligence, School of Computer Science, FEIT, University of Technology Sydney, NSW 2007, Australia
[b] School of Computer Science and Technology, Nantong University, Jiangsu 226019, China
[c] Faculty of Communication Sciences, University of Teramo, I-64100 Teramo, Italy
[d] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Hubei 430074, China

ARTICLE INFO

Article history:
Received 16 May 2019
Received in revised form 30 July 2019
Accepted 3 August 2019
Available online 30 August 2019

Keywords:
Brain–computer interface (BCI)
Clustering
Deep learning
Domain adaptation (DA)
Electroencephalograph (EEG)
Sample selection
Subject adaptation network (SAN)

ABSTRACT

Biosignals tend to display manifest intra- and cross-subject variance, which generates numerous challenges for electroencephalograph (EEG) data analysis. For instance, in the context of classification, the discrepancy between EEG data can make the trained model generalising poorly for new test subjects. In this paper, a subject adaptation network (SAN) inspired by the generative adversarial network (GAN) to mitigate different variances is proposed for analysing EEG data. First the challenges faced by traditional approaches employed for EEG signal processing are emphasised. Then the problem is formulated from mathematical perspective to highlight the key points in resolving such discrepancies. Third, the motivation behind and design principle of the SAN are described in an intuitive manner to reflect its suitability for analysing EEG data. Then after depicting the overall architecture of the SAN, several experiments are used to justify the practicality and efficiency of using the proposed model from different perspectives. For instance, an EEG dataset captured during a stereotypical neurophysiological experiment called the VEP oddball task is utilised to demonstrate the performance improvement achieved by running the SAN.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Research in neuroscience via EEG brain imaging technology has undergone enormous developments over recent years, and the trends are still pushing forward especially when more giant companies are involved [1,2]. The EEG signals captured during the noninvasive process open a window looking into the cognitive process accompanying our daily activities, thus consistently raising the interests of numerous researchers from different perspectives.

Although EEG data can reflect brain activity in some convenient ways, this activity tends to display distinct intra- and cross-subject variations due to individual physiological traits [3]. Such discrepancies pose several challenges for the performance of conventional methods. For example, in some situation the vigilance of the subject is of interest [4], however, EEG signals from the same subject in different runs of the experiment or different subjects with similar vigilance levels tend to present different power spectra. Reciprocally, identical EEG patterns can be labelled quite differently especially from 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 quite different distributions from their counterparts used for prototype method buildup.

Actually, such a problem is not particular for EEG itself. It has been observed and addressed in conventional machine learning studies [5–9], and the corresponding techniques are termed transfer learning (TL) or domain adaptation (DA). However, the difficulties facing EEG lie in the following several aspects: (1) EEG data are immensely complicated and generally of high dimension. Not mention the difficulty to get some intuitive estimations like distributions in different domains, such as training set and testing set. 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 processes. It tends to lead dispute that what kind of features extracted from EEG data suitable for a particular neurological experiment. (3) For some scenario like classification, the ground-truth labels for the corresponding EEG data are sometimes also problematic. For example, in sustained attention tasks [3], the vigilance usually

* Corresponding author.
E-mail addresses: yrming@gmial.com (Y. Ming), ding.wp@ntu.edu.cn (W. Ding), dpelusi@unite.it (D. Pelusi), drwu@hust.edu.cn (D. Wu), Yukai.Wang@uts.edu.au (Y.-K. Wang), Mukesh.Prasad@uts.edu.au (M. Prasad), Chin-Teng.Lin@uts.edu.au (C.-T. Lin).

https://doi.org/10.1016/j.asoc.2019.105689
1568-4946/© 2019 Elsevier B.V. All rights reserved.

indirectly measured via other indicators such as reaction time. However this method unavoidably introduces noise to the labels, which makes applying DA tricks via labels ineffective.

Deep neural networks which are applied in some traditional but difficult fields such as image classification [10] and speech recognition [11], have found success in recent years and achieved state-of-the-art results. There are also innovative ideas like generative adversarial networks (GANs) [12] along the process. Some data in these domains resemble the complexity of EEG data. Therefore, there are attempts to investigate the use of various novel deep network structures and ideas for EEG representation learning, to seek the potentiality of improving EEG analysis as in [13–15].

Work in [16] is regarded as a precursor for adaptation from neural network perspective. It tries to solve the problem with the guidance of GAN and harnessing deep learning's end-to-end modelling philosophy. The work harmonically combines several components together to achieve adaptation. Essentially, the basic idea in [16] is to seek some common representation for both domains. Later, methods such as associative domain adaptation [17] introduce more appropriate loss measurement and generalise the previous work in [16]. Another approach to utilise the neural network for DA is making use of GAN to transform one domain into the other, for example the source domain to the target domain and/or vice versa, as in papers [18,19]. Since the objective is now clearer than the case of finding a common representation space that is not obvious, ADDA and CycleGAN in [18] and [19], respectively, have achieved more astonishing results.

The achievements in [16–19] 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 intra-subject variance, the cross-subject variance has a smaller granularity than the original DA problem. In this paper, 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 adapt EEG data from different experimental sessions of the same subject as well, nevertheless the term subject adaptation network is used instead of session adaptation network to denote its generality. Our contributions lie in several aspects as enumerated:

(1) A theoretical basis is formalised to highlight the key points of adapting EEG data, which not only guides this work based on deep learning, but also can potentially inspire work from other perspectives to address the challenge.

(2) A SAN is designed to mitigate the intra-subject as well as cross-subject variance to facilitate later stage tasks such as sample selection and classification. The designed network draws inspiration from GAN but works in a different way, which opens the possibility of considering GAN in a wider sense to solve other problems.

(3) The research provides illustrative experimental demonstrations of the usefulness and effectiveness of the proposed network architecture, as well as some tricks especially from an implementation perspective.

The rest of the paper is structured as follows. First, the theoretical basis from a mathematical perspective is described to highlight the key points of subject adaptation in Section 2. Based on that the SAN is introduced with the aim of meeting the optimisation constraint in Section 3. In Section 4, the practicality of using the model is demonstrated by running it on several datasets including EEG to justify the innovations from different perspectives. Meanwhile its usage in different analysis contexts are exemplified and the procedures under different requirements summarised. A final discussion is given in Section 5 to emphasise some tricks and pitfalls during our research.

2. Problem description

The EEG signal is the most complex biosignal in physiological research. When EEG signal processing is compared 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 like classification at 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 recently proposed and devised as aforementioned. However, it is still worthwhile to guide the design of the SAN via deductions from the theoretical perspective.

For a given neurological 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 will participate in the experiment several times (or sessions), with each session comprising many experimental trials (or epochs). For analysis, the epoch data for a specific subject, s_i, will be 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. Usually, there is a corresponding label for each trial. Taking the P300 experiment as an example, each trial s_i^j corresponds to a target or a nontarget stimulus [20]. 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 coupled with 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 direct idea is to find another space (called feature space or embedding space) in which the transformed PDFs better align with each other. Defining the mapping from data space to 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 with the following formula:

$$\underset{L}{\operatorname{argmin}} \int \max_i \left\{q_{s_i}(z)\right\} dz \tag{1}$$

$$q_{s_i}(z) = p_{s_i}\left(L^{-1}(z)\right)\left|1/L'\right| \tag{2}$$

An intuitive illustration of (1) is shown in Fig. 1. 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 data space to feature space, the corresponding distributions are denoted by $q_s(z)$ and $q_t(z)$. Suppose some model is trained using s and made the inference on t. For different transformations L_1 and L_2, suppose the learned decision boundaries are those shown in Fig. 1. It is obvious that the performance of generalisation is better in Fig. 1(b) than in Fig. 1(a). The more coherent alignment of the transformed density functions 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, which corresponds to better generalisation.

However, the optimisation of (1) is a variational problem from a mathematical point of view. Since the potential feature space and the form of L are unknown, the problem is generally highly

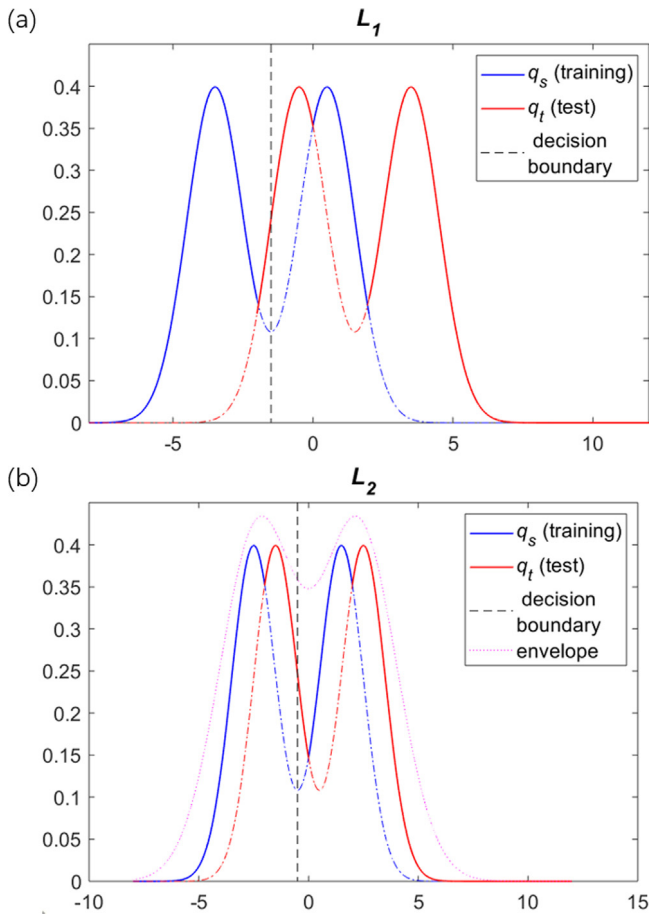

Fig. 1. Illustration of the transformed probability density function alignment. L_2 is preferable to L_1 because the area under the solid line of (b) is smaller than that of (a), which merits the criterion formula (1).

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. 1(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.

3. Proposed subject adaptation network

3.1. Motivation

As mentioned above, the data distribution $p_{s_i}(x, y)$ is usually inconsistent with $p_{s_j}(x, y)$, when subjects participated in the same experiment with an identical setup, or even the same subject participated in the same experiment in multiple runs (now s_i and s_j are referred to 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 optimisation problem does not mean that an analytical solution can be found in practice. For example, a 32-channel EEG data with sampling rate 250 Hz 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 some 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 modality by an adversarial network.

3.2. Constraint realization by neural network

The reasoning and insight above lead to the network architecture designed in Fig. 2. The overall architecture consists of a generative and a discriminative network, resembling the general architecture of GAN [12]. 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 experiment section.

However, our 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.

So, the data flow of our proposed model is as follows: EEG signals either from different sessions or from different subjects are input into the generator for distribution aligning. The dimensionality-reduced latent representations from the generator are pipelined to the discriminator in competition with another discriminator input, which is from 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. 2.

It is noticed that one reason for splitting the generator here into an adaptor plus a mapper, is in the expectation that while the adaptor tries to project the original sample data into some embedding space during the adversarial learning process, it still keeps the projection in reasonable dimensionality. Such a dimensionality is mandatory to have later applications like classification performing 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 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.

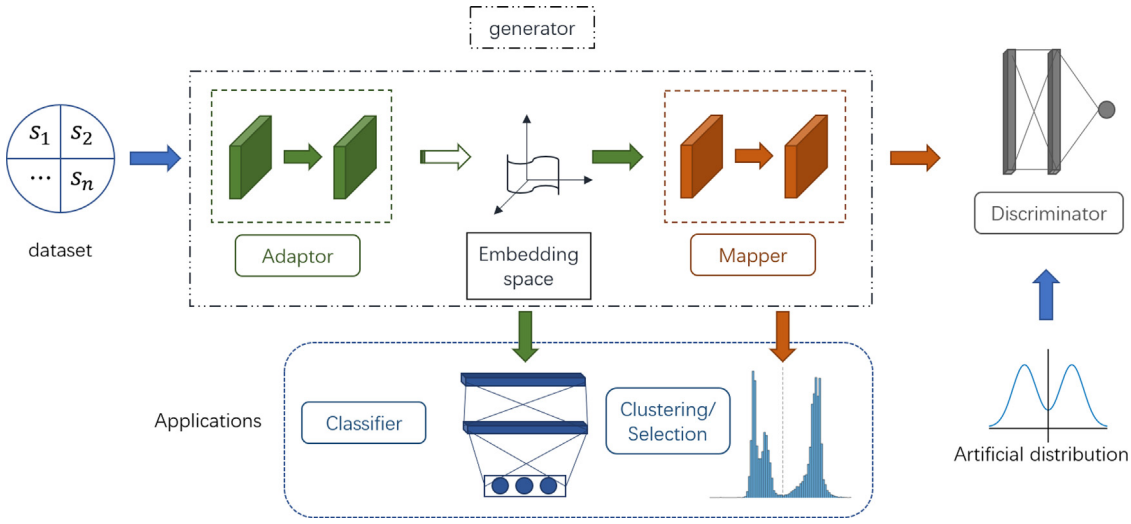

Fig. 2. The overall architecture of the subject adversarial network.

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 aligned coherently and consistently while still keep the modality for later processing such as classification. Consequently, it is clear that the paradigm of utilising our proposed model for EEG data analysis consists of two stages. The first stage is training 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 pipelining the adaptor with or without the mapper to other components according to different applications.

3.3. Artificial target distribution design

Generally, it is impossible to directly observe the distribution of EEG data in 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 [20], 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, just indicated as in Fig. 3.

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.

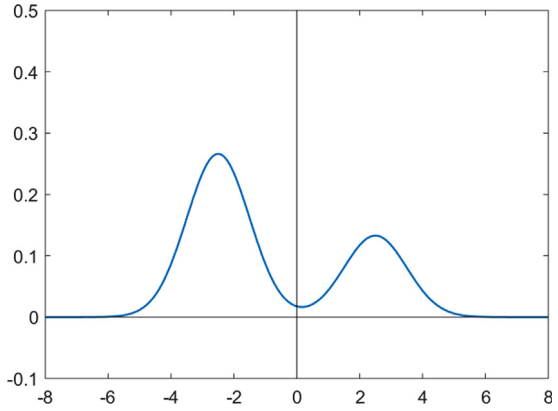

Fig. 3. Designed artificial distribution for an exemplified EEG dataset.

Table 1
Network configuration for MNIST dataset.

Name		Layer	#Filter	Kernel	Activation
Adaptor	L1	Conv2D	32	5×5	–
	L2	Conv2D	64	5×5	–
		AvgPool	–	2×2	–
	L3	Conv2D	128	5×5	–
Mapper	L1	Conv2D	32	3×3	tanh
	L2	Conv2D	64	3×3	tanh
		AvgPool	–	2×2	–
	L3	Linear	256	–	tanh
	L4	Linear	1	–	–
Discriminator	L1	Linear	256	–	ELU
	L2	Linear	64	–	ELU
	L3	Linear	1	–	–

4. Experiment evaluation and analysis

4.1. MNIST dataset

To justify the practicality of using our proposed model, MNIST dataset is utilised for the first stage, namely, data involving a multimodal distribution can be forced to align with the targeted distribution in the low-dimensional space. This process is the basis for later stage applications, and MNIST dataset is chosen here instead of EEG dataset due to its clean label and simplicity especially from the illustrative perspective.

For simplicity, only two handwritten digits, '0' and '1', are filtered out from the original MNIST dataset for demonstration, since it is a little easier to design the bimodal target distribution in this case than using the whole dataset. Considering that the MNIST is a dataset that is nearly balanced over all digits, our designed distribution is in accordance with the following PDF (3).

$$y = 0.5/\sqrt{2\pi} * \exp\left(-(x \pm 2.0)^2/2\right) \tag{3}$$

The configuration of the network is detailed in Table 1. After data preparation, Adam is adopted as the optimiser with batch size 64 and learning rate 1e−4 to train the network for 20 000

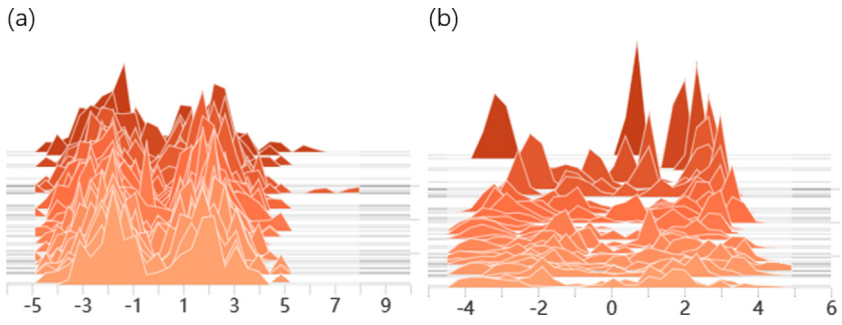

Fig. 4. Histogram illustrations during the training. (a) Histogram of samplings from designed artificial distribution; (b) Histogram of outputs from the generator (adaptor + mapper).

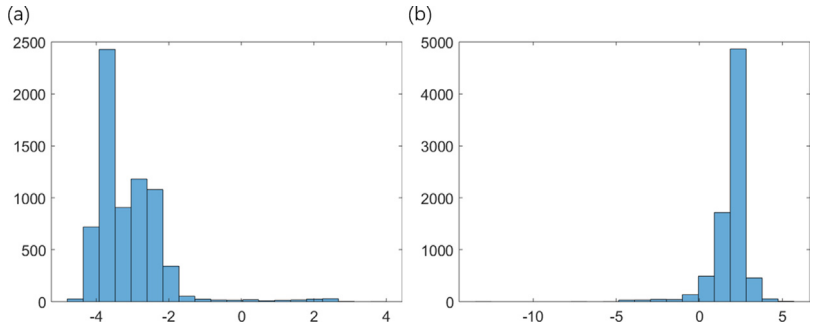

Fig. 5. Histogram of mapped values vs. ground-truth labels (a) Case of digit 0; (b) Case of digit 1.

iterations. To maintain stability during training, the model takes advantage of the loss suggested by Wasserstein GAN (WGAN) [21] as well as spectrum normalisation [22].

Fig. 4 shows the histogram of outputs from the mapper during the training process vs samplings from the designed distribution. It is manifest that such a coherent alignment can be enforced by adversarial training. The ongoing changes with iterations are indicated by the horizontal lines from inwards to outwards. Fig. 4(a) shows the histogram of sampling from the designed artificial distribution, which is coherent during the whole training process. Fig. 4(b) shows the histogram of outputs from the mapper (or the generator). Since the weights are initialised according to some normal distributions, it can be expected to appear as a single modality with a high peak at the beginning. As the training continues, the distribution approaches the targeted distribution. Note that due to the scale problem, the last stages of the histogram in (b) seem to be different from that in (a), but they are actually quite similar.

During the whole process, it is just assumed that handwritten digits of 0 and 1 comply with some bimodal distribution, and no prior information about the labels is utilised. However, after the training process, it is necessary to verify the outputs of the network to justify using our proposed SAN. The histograms of the outputs corresponding to the ground-truth labels are shown in Fig. 5.

It is obvious that the statistical properties of the network outputs are within our expectation. Although the training process assumes no label information, the distributions of each category are clearly separate from each other; nevertheless, there are some outliers. Usually, the variance of 0 is greater than 1 because the strokes of 0 are more complex than those of 1. Consequently, from Fig. 5 it can be observed that the deviation of the 0's distribution is larger than that of the 1's distribution.

Actually, the proposed SAN provides a new alternative to traditional unsupervised clustering methods. For traditional clustering methods such as k-NN, by specifying the number of potential clusters, k-NN relies on the calculation of the distance between

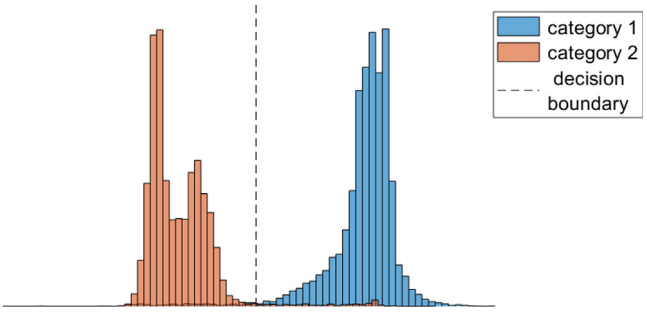

Fig. 6. Clustering according to the distribution of generator outputs.

samples to determine relations. Our method works by approaching the mapping L, which preserves the original distributions. As indicated in Fig. 6, by choosing the decision boundary, a sample x can be assigned to a particular category according to the location of the value $L(x)$ with respect to the decision boundary. Note that in Fig. 6, the plot is against ground-truth labels, however in practice the decision boundary can be empirically drawn since it is quite intuitive to decide in low-dimensional space.

The procedures for performing clustering based on the SAN are summarised as follows:

Estimate the statistical property of the sample data $\{x_i\}$.
Design the artificial probability density function f.
Train the network to obtain the optimised mapping function $L(x;\theta^*)$ (generator network).
Observe the histogram of network outputs hist $(L(\{x_i\};\theta^*))$ which has a low dimensionality such as 1D or 2D.
Decide where the decision boundary B is located.
Decide the category of a given sample x' by comparing $L(x')$ and B.

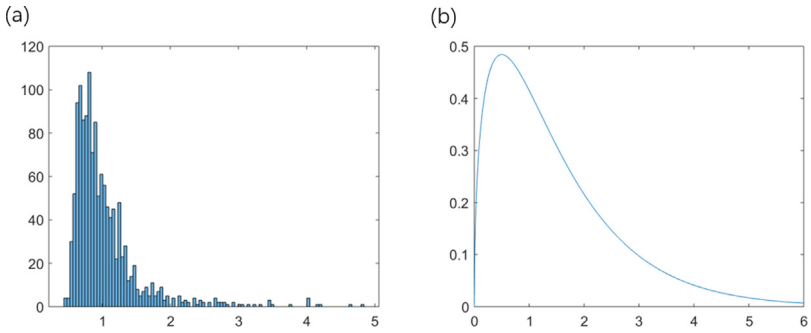

Fig. 7. (a) Histogram of measured RTs; (b) Designed gamma distribution approximating the distribution of measured RT.

4.2. Driving EEG dataset

As a brain imaging technology, EEG can reflect some characteristics of mind state or cognitive process pertaining to people's specific activities like driving. By collectively analysing samples from test subjects, some conclusions especially from neurophysiological aspect can be reached, such as activities of different lobes, connectivity between lobes, etc. However, the difficulty in EEG research is by estimating such implicit mind states to select the most appropriate samples for analysis. For example, to understand the brain dynamic under low performance driving, one needs to screen out the trial data corresponding to fatigue for analysis.

However, the mind states or attention levels are by no means can be directly measured. It relies on a clever experiment design to indirectly measure such indicators. For example, in simulated driving experiment, the reaction time (RT) is utilised for labelling attention level. But usually the relation between the direct but implicit label such as attention level with indirect but explicit indicator like RT is unknown. And due to other factors, such measurement is usually error-prone or at least accompanied by noise. It can potentially move the EEG data actually corresponding to the alert stage into the fatigue category, consequently it will hurt the accuracy of the conclusions. Actually, even for one subject, the higher number of runs the subject participate in the experiment, the more likely of inconsistency on data labelling. Now and then, such intra-subject variance poses a challenge for effective analysis.

In this experiment, a driving EEG dataset is utilised to demonstrate the capability of our proposed network for sample selection, especially from the intra-subject variance perspective. The EEG data were captured during a simulated sustained-attention driving task which is to investigate people's performance during driving under different vigilance levels. The setup is to have test subjects driving in a simulated four-lane highway for an enduring test lasting 90 min. It is based on the postulate that the attention of test subjects cannot be maintained at the same level during the entire procedure, which results in reactions to occurrences requiring instant reactions during driving exhibiting various latencies.

To measure the subjects' vigilance during driving, lane-departure events were deliberately and randomly introduced by having the car drift away from the original cruising lane towards the left or right side (deviation onset). Test subjects were instructed to quickly compensate for this perturbation by steering the wheel (response onset), to turn the car back to the original lane (response offset). It is manifest that the extent of fatigue is closely related to the latency of response onset. The duration between such consecutive event onsets, called reaction time (RT), is utilised to label the EEG data into different vigilance stages. The detailed explanations of the baseline period and consecutive events for each complete trial can be found in [23].

Table 2
Network configuration for driving dataset.

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

EEG signals are recorded simultaneously and continuously using the SynAmps2 Express system during the experiment. To increase the correlation between fatigue and RT as well as to exclude other impact factors during testing, participants need to operate only the steering wheel in reacting to lane-perturbation events and are free from controlling the accelerator and brakes. However, due to the inclination of subjects to be distracted, some RT cannot faithfully reflect the underlying fatigue state.

As mentioned above, one subject who participated in the experiment several times with the greatest number of trials is chosen to demonstrate utilising the model for solving the intra-subject variance problems aka sample selection. For emphasising the key steps, just the alpha band (4∼7 Hz) of the EEG data is chosen. It is converted to topographies after pre-processing (down-sampling from its original 500 Hz to 250 Hz followed by a band-pass filter with range 0.5 to 50 Hz) with corresponding RTs.

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. 7(a) is the histogram of RT, hence a gamma distribution is chosen in Fig. 7(b) as the designed distribution.

With the specified configuration in Table 2, the EEG data are processed in spatial domain by training the network for 4e+4 iterations with a learning rate 5e−4 until the enforced alignment aka the gamma distribution is clearly observed. The enforced distribution of samples vs corresponding RTs is plotted in Fig. 8 to select the most appropriate samples for further analysis. In Fig. 8, according to our domain knowledge, the orange dashed eclipse is plotted to indicate our recommendation of EEG samples for corresponding analysis, since the measured RTs are more consistent with the enforced aligned distribution.

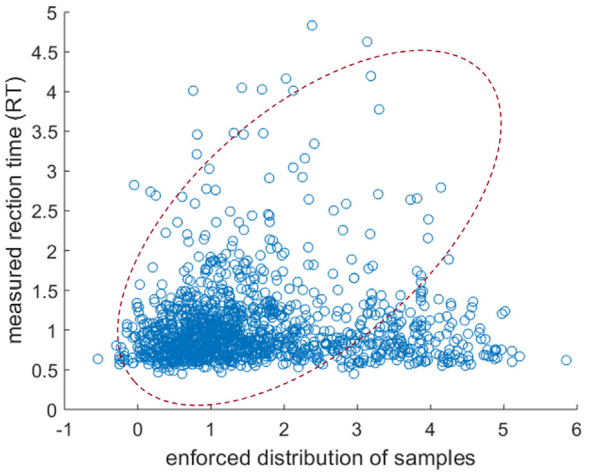

Fig. 8. Plotting of enforced sample distribution vs. measured RT. The dashed eclipse indicates the recommended range for selecting samples.

To demonstrate such a claim, the topographies of neighbouring samples in three groups are inspected along the line $y = 1.5$ in Fig. 9. 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 our claims as well as the suitability of utilising our 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 optimised 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. 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 a visually evoked potential (VEP) oddball task [20] is utilised here. The experiment is based on the P300 (P3), an event-related potential (ERP) elicited in the visual system. During the experiment, image stimuli are presented to subjects at a rate of 0.5 Hz (one image every two seconds). The dichotomous images were either an enemy combatant or a U.S. soldier. The subjects were instructed to identify each image as a target or a nontarget with a unique button press as quickly but as accurately as possible.

The image set used in this experiment is an imbalanced version with 34 targets among 270 images. Eighteen subjects participated in the experiments, which lasted for 15 min on average. In this work, signals recorded with a wired 64-channel ActiveTwo3 system (sample rate set to 512 Hz) from BioSemi is used for data acquisition. More details especially the consent of the experiment can be found in [20].

EEG data is usually processed in the frequency domain [24]. By sacrificing time resolution, it is expected the frequencies or power spectrum can unveil more statistical properties of underlying brain activities in a certain period. Usually, data analysis requires using domain knowledge to evaluate the viability of the feature to be used or the way to extract features. However, due to the limited knowledge of the cognitive or physiological process, it is still not very clear which are the best sub-band or novel features buried in EEG data or even whether the frequency domain is the best domain in which to work or not. Considering the feature extraction capabilities of deep neural networks, it is persuasive to directly work with waveform EEG data in the time domain. Works in [15] and [25] launched this aspect of research and achieved some promising results, so this paper takes a similar approach.

Table 3
Network configuration for Oddball dataset.

Name		Layer	#Filter	Kernel	Activation
Adaptor	L1	Conv2D	32	7×7	–
		AvgPool	–	2×2	–
	L2	Conv2D	64	5×5	–
		AvgPool	–	2×2	–
	L3	Conv2D	128	3×3	–
		AvgPool	–	2×2	–
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	–	–

Usually, EEG data tend to undergo a series of pre-processing such as bandpass filtering, down-sampling before feeding into the network, meantime to retain as much information as possible to allow the network to automatically discover the most appropriate latent representations. For data preparation in this paper, 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 normalisation. To counter the potential bias during training, the data are resampled to have a ratio of targets to nontargets as 1:3, which is also taken into consideration when the artificial distribution is designed.

Fig. 2 suggests that during adaptation, the outputs of some intermediate layer can be to some extent aligned more coherently than the data in the original sample space, while they still have reasonable dimensionality for classification. This point of view is highlighted by the embedding space and its relationship with the classifier in Fig. 2. It is mentioned that there are 382 targets vs. 1146 nontargets. The dimension of sample space is 2880 (64 channels with each channel having 45 data points). Such a limited number of samples in such a high-dimensional space are far from enough to capture the underlying data distribution. The available data constraint leaves us with the choice to practically analyse only one single channel which is Pz, to restrict the sample space to 45 dimensions. Notice the data can still be treated from the two-dimensional perspective just with the height equal to 1 here. Based on this limitation, a network configuration is given in Table 3.

The designed artificial distribution complies with the following Eq. (4):

$$y = (0.5 \mp 0.25) / \sqrt{2\pi} * \exp\left(-(x \pm 2.0)^2 / 2\right) \tag{4}$$

With the targeted distribution, the network is trained with the Adam optimiser of learning rate 1e−4 for 1e+4 iterations. The enforced distribution with ground-truth labels is shown in Fig. 10. Due to the complexity of EEG data, the separation is not as clear as the case of MNIST data.

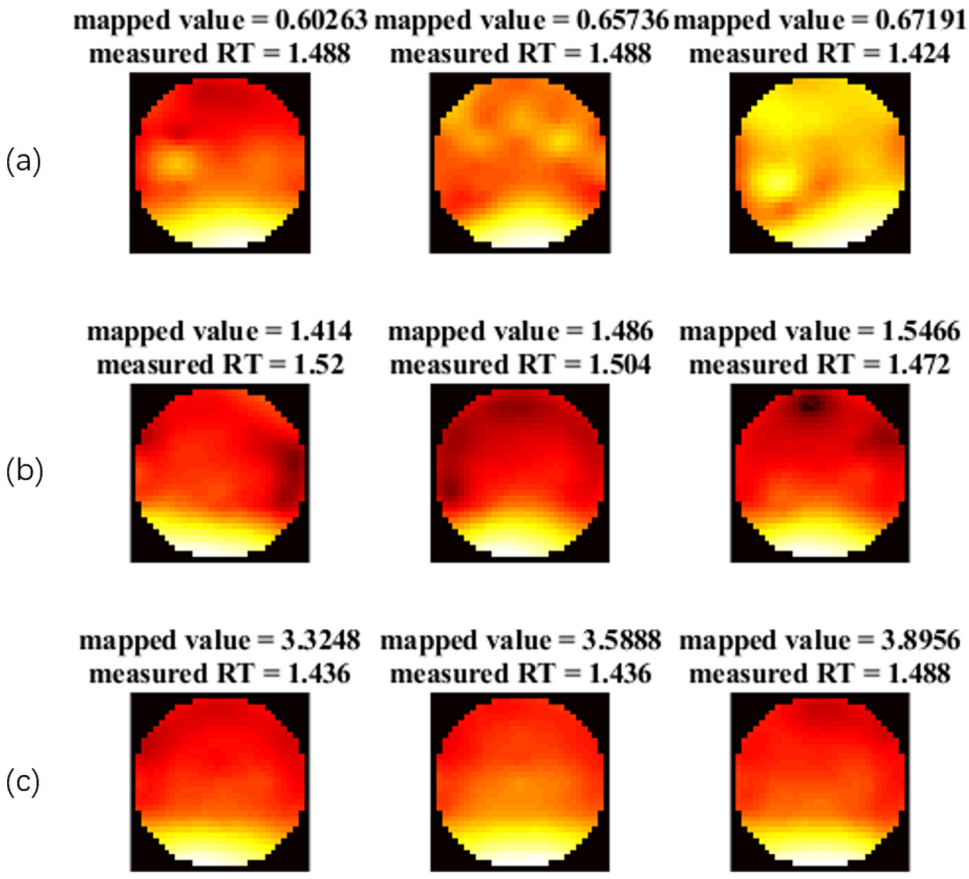

Fig. 9. Topographies of the samples in different groups. (a), (b) and (c) are chosen based on the measured RT and mapped values of the enforced distribution.

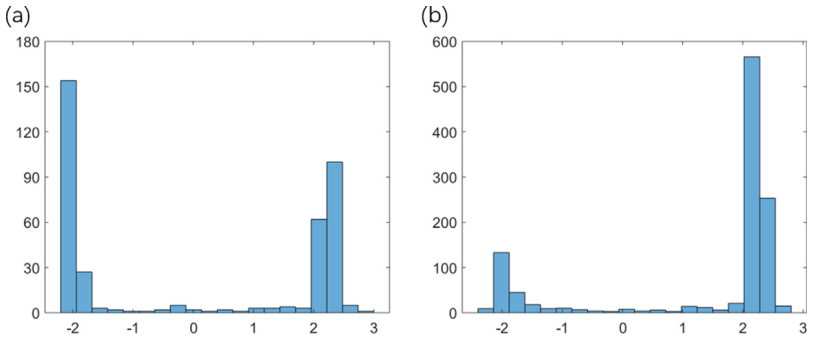

Fig. 10. (a) Histogram of samples corresponding to target (b) Histogram of samples corresponding to non-target.

Table 4
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**

For benchmarking comparisons, the conventional support vector machine (SVM) method and the current well-spoken-of EEGNet [15] are chosen here. Since the data is just of single-channel, the depthwise convolution which is analogous to the common spatial pattern (CSP) filter [26] is omitted from the original EEGNet structure. For the sake of brevity, only the first five subjects are utilised for leave-out testing.

The results are listed in Table 4 for comparison. With budget EEG data, all the models produce comparable results, but our proposed model is slightly better among all. It is mentioned that SVM has a big advantage with limited training samples, because only a few support vectors can help determine the decision boundary with reasonable margin. However, it is believed that by increasing the number of samples to a reasonable magnitude which is appropriate to train our proposed network in the adversarial way, the improvement could be further boosted. Nevertheless, the improvement of subjects' adaptation can be justified from the results, even with limited samples.

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

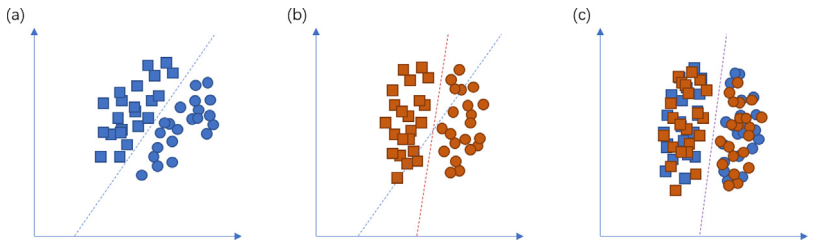

Fig. 11. (a) Distribution of training samples and the corresponding decision boundary; (b) Distribution of test samples and the corresponding decision boundary as well as migrated decision boundary from training; (c) Adapted training samples and test samples and corresponding decision boundary. Note the coordinate system in (c) is not necessarily the same as (a) and (b), because the adapted samples can be in some latent space.

Estimate the sample distribution.
Design the artificial probability density function f that simulates the estimated sample distribution.
Train the network to obtain the optimised 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.

5. Discussion

The first topic for discussion is about the motivations which drive this research. To balance between the brevity and plain explanation for some background information of this work, it has been moved to this section. Due to the long-standing research of domain adaptation, one motivation is just recapped here to highlight its necessity. As in Fig. 11, suppose a model is constructed to target a binary classification problem. In Fig. 11(a), a training process leads to a decision boundary be nicely set up between the two training class 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. 11(b). Since it is presumed that training set and test set comply with some bimodal distributions, if 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. 11(c).

This is exactly the case of EEG data, where data for training and testing respectively can be quite different from each other. Such a cross-subject variance problem just inherits the common problem of domain adaptation which tries to address, and is worthy of research as exemplified in this work. However, besides that, EEG data is endowed with another trait, aka intra-subject variance, as the case demonstrated in Fig. 12. It shows the RTs of one subject participated in the experiment for two times. According to the experiment log, these two experiments were carried out with identical setup, similar time, and subject reported similar physical and mental conditions when participating. Nevertheless, the eminent discrepancy between RT patterns for different sessions of the same subject posts a big challenge for selecting the appropriate EEG data segments to analyse the neurophysiological process during the experiment.

Hence, to select the most appropriate samples to analyse the underlying cognitive process like attention alteration, it requires not only the measure RT being considered, but also the maximum commonality between sessions being sought. For example, during which stage of the experiment the subject is anticipated as alertness, which stage the subject is as drowsiness. This requires some adaptation method to sort out the samples coherent with the distribution of RTs, another motivation for our research. As shown in Fig. 8, the samples which distribute along the diagonal line are preferred.

Next some tricks that are recognised by us as subtle and useful are highlighted to help ease 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 since the details are blurred to some extent. One consequence is the potentially induced higher variance as shown in Fig. 13. Another consequence is the potential failure of adaptation when the variance exceeds a certain width, which means the alignment cannot be effectively enforced.

It is obvious 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 hyperbolic tangent function is chosen because the target distribution that being designed is to some extent symmetric with respect to $x = 0$. For the second experiment the sigmoid function is chosen since gamma distribution only exists when $x > 0$. It is also noticed that the mapper is quite sensitive to initialisation. When Gaussian normalisation is used, the standard deviation is suggested to keep within 0.5. Using larger values tend to cause modal collapse. One possible reason is the initial penalty is likely so severe that it restricts the exploration throughout the whole training process. This comparison is demonstrated in Fig. 14.

It is also found that it is a little 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 have been 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 our 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 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 our 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 utilise 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 in the future.

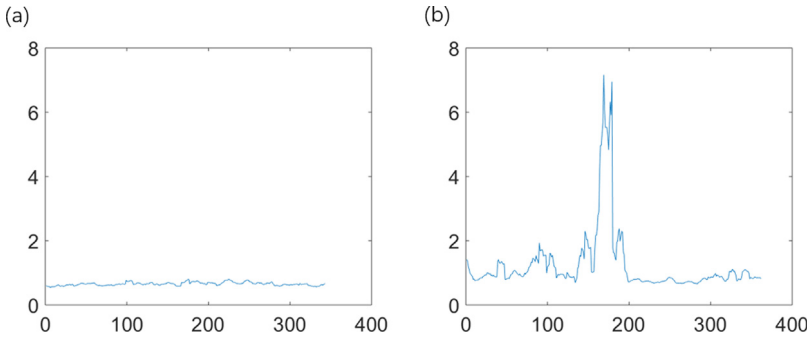

Fig. 12. RT patterns for the same subject who participated in the experiment for two times.

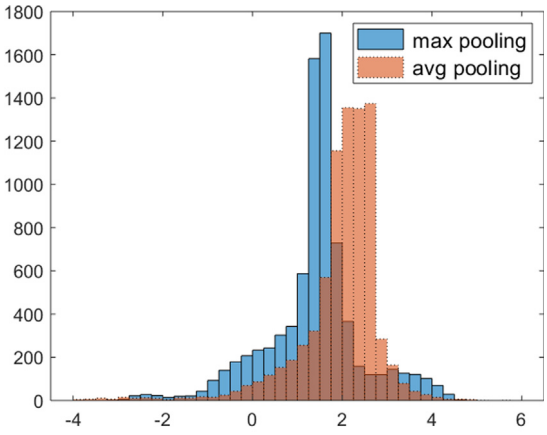

Fig. 13. The impact of different pooling methods on the enforced distribution of samples.

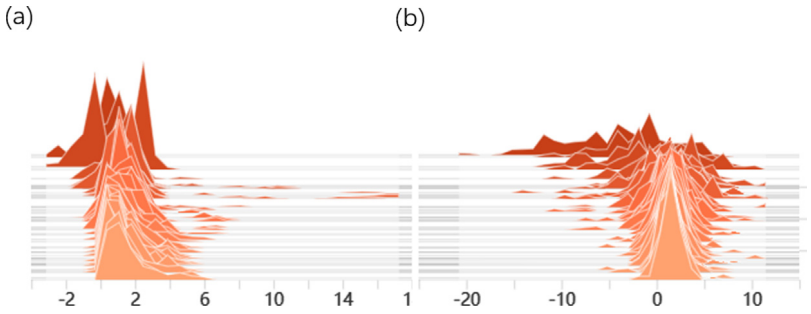

Fig. 14. Enforced distribution trends during training with different standard deviation for the second experiment (a) $\sigma = 0.2$ (b) $\sigma = 1.0$.

6. Conclusion

In this paper a subject adversarial network is proposed with the guidance of a variational optimisation problem targeting the mitigation of the intra- and cross-subject variance among EEG data. The architecture and usage of the network are detailed and demonstrated from different perspectives with respect to EEG data analysis, such as sample selection and classification. The experimental results indicate its practicality and efficiency in different scenarios. Some tricks developed during the research are also discussed with goodwill of their potential helpfulness concerning adversarial training for further applications in wider sense. Meanwhile, the effectiveness of utilising our proposed model depends on the designed artificial distribution, which demands the understanding of the problem domain and inspection of the data to be analysed. These constraints might require extra work compared with straight-forward usage of other models. Future work can be considered from potential auxiliary methods for better distribution designed and increase of data scale which is critical for the training of deep neural network models.

Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.asoc.2019.105689.

Acknowledgements

This work was supported in part by the Australian Research Council (ARC) under discovery grant DP180100670 and DP180100656; The National Natural Science Foundation of China under Grant 61976120; The Natural Science Foundation of Jiangsu Province under Grant BK20151274; The Qing Lan Project of Jiangsu Province; NSW Defence Innovation Network and NSW

State Government of Australia under the grant DINPP2019 S1-03/09; Office of Naval Research Global, US under Cooperative Agreement Number ONRG-NICOP-N62909-19-1-2058.

References

[1] Eliza Strickland, Facebook Announces Typing-by-Brain Project, Spectrum IEEE https://spectrum.ieee.org/the-human-os/biomedical/bionics/facebook-announces-typing-by-brain-project, 2017.

[2] NEUROLINK. http://www.neurolinkglobal.com, 2018.

[3] Dongrui Wu, Chun-Hsiang Chuang, Chin-Teng Lin, Online driver's drowsiness estimation using domain adaptation with model fusion, in: International Conference on Affective Computing and Intelligent Interaction (ACII), 2015.

[4] Chin-Teng Lin, Chun-Hsiang Chuang, Chih-Sheng Huang, Yen-Hsuan Chen, Li-Wei Ko, Real-time assessment of vigilance level using an innovative mindo4 wireless EEG system, in: IEEE International Symposium on Circuits and Systems (ISCAS), 2013.

[5] J. Jiang, A literature survey on domain adaptation of statistical classifiers. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey, (2008).

[6] M. Sugiyama, T. Suzuki, et al., Direct importance estimation for covariate shift adaptation, Ann. Inst. Statist. Math. 60 (2008) 699–746.

[7] B. Zadrozny, Learning and evaluating classifiers under sample selection bias, in: Proc. ICML, 2004, pp. 114–121.

[8] Kun Zhang, Bernhard Scholkopf, Krikamol Muandet, Zhikun Wang, Domain adaptation under target and conditional shift, in: International Conference on Machine Learning, 2013, pp. 819–827.

[9] Kate Saenko, Brian Kulis, Mario Fritz, Trevor Darrell, Transferring Visual Category Models to New Domains, Technical Report No. UCB/EECS-2010-54 (2010).

[10] Krizhevsky Alex, Sutskever Ilya, E. Hinton Geoffrey, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.

[11] A. Graves, A.R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial nets, Adv. Neural Inf. Process. Syst. (2014) 2672–2680.

[13] Nan Zhang, Wei-Long Zheng, Wei Liu, Bao-Liang Lu, Continuous vigilance estimation using LSTM neural network, in: ICONIP, 2016.

[14] Pouya Bashivan, Irina Rish, Mohammed Yeasin, Noel Codella, Learning representations from EEG with deep recurrent-convolutional neural network, in: ICLR, 2016.

[15] Vernon J. Lawhern, Amelia J. Solon, et al., EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces, arXiv:1611.08024v3, 2018.

[16] Yaroslav Ganin, Victor Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 1180–1189.

[17] Philip Haeusser, Thomas Frerix, Alexander Mordvintsev, Daniel Cremers, Associative Domain Adaptation, arXiv:1708.00938v1 [cs.CV] (2017).

[18] Eric Tzeng, Judy Hoffman, Kate Saenko, Trevor Darrell, Adversarial Discriminative Domain Adaptation, arXiv:1702.05464v1 [cs.CV] (2017).

[19] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, arXiv:1703.10593v5 [cs.CV] (2018).

[20] Dongrui Wu, Vernon J. Lawhern, W. David Hairston, Brent J. Lance, Switching EEG headsets made easy: Reducing offline calibration effort using active weighted adaptation regularization, IEEE Trans. Neural Syst. Rehabil. Eng. 24 (11) (2016) 1125–1137.

[21] Martin Arjovsky, Soumith Chintala, Léon Bottou, G.A.N. Wasserstein, arXiv:1701.07875v3 (2017).

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, Yuichi Yoshida, Spectral Normalization for Generative Adversarial Networks, arXiv:1802.05957v1 (2018).

[23] Yurui Ming, Yu-Kai Wang, Mukesh Prasad, Dongrui Wu, Chin-Teng Lin, Sustained attention driving task analysis based on recurrent residual neural network using EEG data, in: IEEE International Conference on Fuzzy Systems, 2018.

[24] Sanei Saeid, EEG Signal Processing, John Wiley & Sons, 2007.

[25] Schirrmeister, Robin Tibor, et al., Deep learning with convolutional neural networks for EEG decoding and visualization, Hum. Brain Mapp. 38 (11) (2017) 5391–5420.

[26] K.K. Ang, Z.Y. Chin, C. Wang, C. Guan, H. Zhang, Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b, Front. Neurosci. 6 (2012) 39.