

SYSTEM AND METHOD FOR UNSUPERVISED FEW-SHOT ADAPTIVE RE-LEARNING OF A DEEP LEARNING NEURAL NETWORK

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of priority to the Singapore application no. 10202260088P filed on 14 November 2022, the contents of which are hereby incorporated by reference in their entirety for all purposes.

TECHNICAL FIELD

[0002] This application relates to a system and method for performing targeted model re-learning of a trained deep learning neural network based on a small sample of data obtained from unlabeled evaluation data.

BACKGROUND

[0003] To address the challenge of limited subject-specific training data available for electroencephalogram (EEG) motor-imagery classification, conventional subject-independent machine learning approaches have been employed. The methods proposed in the art typically involve the co-training of machine learning models with data from other subjects. However, it is essential to acknowledge that considerable signal variability exists between subject-to-subject and session-to-session, and this introduces noise and bias into the training process, ultimately negatively affecting model performance. This is evident in some cases where the predictions suffer from decreased accuracy following the re-adaptation of the trained model, underscoring the complexity of controlling adaptive performance.

[0004] Furthermore, while it has been noted that performances of trained neural networks have been enhanced through a fine-tuning process, such an approach often necessitates a substantial volume of target data to yield noticeable improvements and mitigate overfitting. Additionally, existing fine-tuning techniques mandate the availability of labeled data, rendering such algorithms resource-intensive and unfeasible for real-time, online applications.

[0005] A limitation inherent to approaches proposed by those skilled in the art lies in the system's inability to selectively control the data that is utilized in training the model. Consequently, these methodologies may lack robustness when faced with anomalous data or when the underlying dataset contains complex and ambiguously defined latent features. Hence,

those skilled in the art are constantly looking for ways to fine-tune a trained neural network using an unsupervised few-shot deep representation learning technique.

SUMMARY

[0006] In one aspect, the present application discloses a computer system configured to fine-tune a trained neural network. The computer system comprises a computer processor; and a non-transitory media readable by the processing unit, the media storing instructions that when executed by the processing unit, causes the processing unit to predict, using the trained neural network, latent features from a first plurality of unlabeled non-target data, whereby the neural network was trained based on a second plurality of unlabeled non-target data and predict, using the trained neural network, latent features of a plurality of labeled target data. The processing unit is then directed to compute similarity values between the predicted latent features of the first plurality of unlabeled non-target data and the predicted latent features of the plurality of labeled target data and to then select unlabeled non-target data that have similarity values below a defined threshold from the first set of unlabeled non-target data; and to fine-tune the trained neural network by re-training the trained neural network using the selected unlabeled non-target data.

[0007] In another aspect, the present application discloses a method to fine-tune a trained neural network using a computer system, the method comprising the steps of predicting, using the trained neural network, latent features from a first plurality of unlabeled non-target data, whereby the neural network was trained based on a second plurality of unlabeled non-target data; predicting, using the trained neural network, latent features of a plurality of labeled target data; computing, using the computer system, similarity values between the predicted latent features of the first plurality of unlabeled non-target data and the predicted latent features of the plurality of labeled target data; selecting, using the computer system, unlabeled non-target data that have similarity values below a defined threshold from the first set of unlabeled non-target data; and fine-tuning, using the computer system, the trained neural network by re-training the trained neural network using the selected unlabeled non-target data.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] Various embodiments of the present disclosure are described below with reference to the following drawings:

Figure 1 illustrates a block diagram of a Siamese-variational autoencoder (VAE) unsupervised few-shot adaptive framework in accordance with embodiments of the present disclosure;

Figure 2 illustrates a block diagram representative of a processing system for performing embodiments of the present disclosure;

Figure 3 illustrates a block diagram of the modules in a proposed VAE network architecture in accordance with embodiments of the present disclosure;

Figures 4a and 4b illustrate exemplary structuring of target data in accordance with embodiments of the present disclosure;

Figures 5a and 5b illustrate a comparison of individual subjects' performance between a state-of-the-art framework and the system proposed in Figure 1 in accordance with embodiments of the disclosure;

Figures 6a and 6b illustrate a t-Distributed Stochastic Neighbor Embedding (t-SNE) plots of learned representations from a subject;

Figure 7 illustrates a plot of feature values against time of learning representations from a subject for an online setting and an offline setting;

Figure 8 illustrates a plot of a Savitzky-Golay moving average of the Euclidean distance between the validation and evaluation trials against the trials in the transfer set; and

Figure 9 illustrates a flowchart of a process for fine-tuning a trained neural network in accordance with embodiments of the present disclosure.

DETAILED DESCRIPTION

[0009] The following detailed description is made with reference to the accompanying drawings, showing details and embodiments of the present disclosure for the purposes of illustration. Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments, even if not explicitly described in these other embodiments. Additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.

[0010] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements.

[0011] In the context of various embodiments, the term “about” or “approximately” as applied to a numeric value encompasses the exact value and a reasonable variance as generally understood in the relevant technical field, e.g., within 10% of the specified value.

[0012] As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0013] As used herein, “comprising” means including, but not limited to, whatever follows the word “comprising”. Thus, use of the term “comprising” indicates that the listed elements are required or mandatory, but that other elements are optional and may or may not be present.

[0014] As used herein, “consisting of” means including, and limited to, whatever follows the phrase “consisting of”. Thus, use of the phrase “consisting of” indicates that the listed elements are required or mandatory, and that no other elements may be present.

[0015] As used herein, the terms “EEG signal” or “EEG data” generally refer to Electroencephalography (EEG) signals or EOG Electrooculograms (EOG) signals, or a combination thereof, obtained from one or more subjects whereby these signals correspond to the subject’s brain activity.

[0016] As used herein, “target data” refers to data obtained from a particular or a specific test subject(s), and/or subject-specific training data for electroencephalogram (EEG) motor-imagery classification while “non-target data” refers to data obtained from a group of randomly selected subjects and/or non-subject specific training data for EEG motor-imagery classification.

[0017] Further, one skilled in the art will recognize that certain functional units in this description have been labelled as modules throughout the specification. The person skilled in the art will also recognize that a module may be implemented as circuits, logic chips or any sort of discrete component. Furthermore, one skilled in the art will also recognize that a module may be implemented in software which may then be executed by a variety of processor architectures. In embodiments of the disclosure, a module may also comprise computer instructions or executable code that may instruct a computer processor to carry out a sequence

of events based on instructions received. The choice of the implementation of the modules is left as a design choice to a person skilled in the art and does not limit the scope of the claimed subject matter in any way.

[0018] Further, one skilled in the art will also recognize that the detailed workings, internal structures, techniques and training datasets of machine learning models, neural networks and anomaly detection models referred to in this disclosure have not been disclosed or shown in detail as such information is known to one skilled in the art and as such, has been omitted for brevity.

[0019] It is useful to note that electroencephalography (EEG) is a commonly employed method for directly measuring human brain activity. It involves the placement of multiple electrodes on the surface of the human head to collect EEG signals. With its high temporal resolution, EEG can capture dynamic changes in brain activity at a sub-second level.

[0020] Variational autoencoders (VAEs) comprise of probabilistic encoder-decoder pairs. For a given input x , the encoder, which is an inference model with weights and biases Θ , will map the input x to a lower-dimensional space, which is referred to as the latent space. This mapping results in the generation of a set of hidden latent variables and these variables may be defined as output z . In embodiments of the disclosure, the inference model may be defined by a Gaussian probability distribution $q_{\Theta}(z|x)$. For the same VAE, a decoder model with weights and biases represented by φ may be represented by a joint probability function, $\rho_{\varphi}(x, z) = \rho_{\varphi}(x|z)\rho(z)$. During the training of the VAE, the encoder and decoder are both trained simultaneously with the goal of finding the parameters that best optimize the variational lower bound of the likelihood of the observed input data, $\rho_{\varphi}(x) = \int \rho_{\varphi}(x, z)dz$ while ensuring that the latent space z adheres to a prior Gaussian distribution.

[0021] Thus, the effectiveness of the VAE in reconstructing the original input is given by the reconstruction log-likelihood function, $\log \rho_{\varphi}(x|z)$. The reconstruction loss function, $-EQ_{\Theta}(z|x)[\log \rho_{\varphi}(x|z)]$, is therefore given by the expected negative log-likelihood function computed with respect to the distribution of the latent features by the encoder, where $EQ_{\Theta}(z|x)$ is defined as the expected value taken with respect to the distribution of latent features (z) given input data (x) and $-\log \rho_{\varphi}(x|z)$ is defined as the negative log-likelihood function.

[0022] In addition to the reconstruction loss, VAEs also take into account a regularisation term given by the Kullback-Leibler (KL) divergence between two continuous distributions - the encoder's variational posterior $q_{\theta}(z|x)$ and the prior distribution $\rho(z)$ for the latent variables. The KL divergence measures how close the two distributions $q_{\theta}(z|x)$ and $\rho(z)$ are to each other and the KL divergence function is defined as $KL(q_{\theta}(z|x)||\rho(z))$. Therefore, based on the above, the overall loss function L_i for an input datapoint x_i may be defined as:

$$L_i = -EQ_{\theta}(z|x_i)[\log \rho_{\phi}(x_i|z)] + KL(q_{\theta}(z|x_i)||\rho(z)) \quad \dots(1)$$

where the KL divergence is computed by:

$$KL(q_{\theta}(z|x_i)||\rho(z)) = \sum_{x \in \mathcal{X}} q_{\theta}(z|x_i) \log \frac{q_{\theta}(z|x_i)}{\rho(z)} \quad \dots(2)$$

[0023] Figure 1 illustrates a block diagram of a Siamese-neural network unsupervised few-shot adaptive framework that may be implemented in computer system 100 in accordance with embodiments of the present disclosure. In embodiments of the disclosure, the framework illustrated in Figure 1 comprises a Siamese- variational autoencoder (VAE) framework. During a training phase, a neural network such as, but is not limited to, a VAE model (not shown) may be trained based on a plurality of unlabeled non-target data. Once the training process is completed, trained VAE model 104 may then be applied to unlabeled non-target data 101, e.g., $Y_n = 1 \dots N$, to produce predicted output 103, e.g., $\tilde{Y}_n = 1 \dots N$.

[0024] Trained VAE module 104 may also then be applied to labeled target data 105, e.g., X_n , to obtain predicted dataset 107, e.g., $\tilde{X}_n$. During the generation of predicted output 103, a set of latent features z_n will be computed and during the generation of predicted output 107, a set of latent features z_x will similarly be computed. It should be noted that data from labeled target data 105 does not overlap with unlabeled non-target data 101.

[0025] At stage 108, computer system 100 then proceeds to compute similarity values between the latent features z_n and the latent features z_x . In embodiments of the disclosure, this computation is done based on the contrastive distances between the two sets of latent features. Unlabeled non-target data having similarity values below a defined threshold of being the 90th

percentile and below in terms of contrastive distance from the target subject are then selected from unlabeled non-target data 101 by computer system 100 at stage 109, and this selected unlabeled non-target data is then used by computer system 100 to fine-tune trained VAE model 104 at stage 110.

[0026] In accordance with embodiments of the present disclosure, a block diagram representative of components of processing system 200 that may be provided within computer system 100 to carry out any of the functions described above is illustrated in Figure 2. One skilled in the art will recognize that the exact configuration of each processing system provided within these modules may be different and the exact configuration of processing system 200 may vary and the arrangement illustrated in Figure 2 is provided by way of example only.

[0027] In embodiments of the disclosure, processing system 200 may comprise controller 201 and user interface 202. User interface 202 is arranged to enable manual interactions between a user and the computing module as required and for this purpose includes the input/output components required for the user to enter instructions to provide updates to each of these modules. A person skilled in the art will recognize that components of user interface 202 may vary from embodiment to embodiment but will typically include one or more of display 240, keyboard 235 and optical device 236.

[0028] Controller 201 is in data communication with user interface 202 via bus 215 and includes memory 220, processor 205 mounted on a circuit board that processes instructions and data for performing the method of this embodiment, an operating system 206, an input/output (I/O) interface 230 for communicating with user interface 202 and a communications interface, in this embodiment in the form of a network card 250. Network card 250 may, for example, be utilized to send data from these modules via a wired or wireless network to other processing devices or to receive data via the wired or wireless network. Wireless networks that may be utilized by network card 250 include, but are not limited to, Wireless-Fidelity (Wi-Fi), Bluetooth, Near Field Communication (NFC), cellular networks, satellite networks, telecommunication networks, Wide Area Networks (WAN) and etc.

[0029] Memory 220 and operating system 206 are in data communication with CPU 205 via bus 210. The memory components include both volatile and non-volatile memory and more

than one of each type of memory, including Random Access Memory (RAM) 223, Read Only Memory (ROM) 225 and a mass storage device 245, the last comprising one or more solid-state drives (SSDs). One skilled in the art will recognize that the memory components described above comprise non-transitory computer-readable media and shall be taken to comprise all computer-readable media except for a transitory, propagating signal. Typically, the instructions are stored as program code in the memory components but can also be hardwired. Memory 220 may include a kernel and/or programming modules such as a software application that may be stored in either volatile or non-volatile memory.

[0030] Herein the term “processor” is used to refer generically to any device or component that can process such instructions and may include: a microprocessor, microcontroller, programmable logic device or other computational device. That is, processor 205 may be provided by any suitable logic circuitry for receiving inputs, processing them in accordance with instructions stored in memory and generating outputs (for example to the memory components or on display 240). In this embodiment, processor 205 may be a single core or multi-core processor with memory addressable space. In one example, processor 205 may be multi-core, comprising—for example—an 8 core CPU. In another example, it could be a cluster of CPU cores operating in parallel to accelerate computations.

[0031] In accordance with embodiments of the disclosure, the trained neural network may comprise a trained VAE model 104 which in turn has the network architecture illustrated in Figure 3. In particular, the network architecture comprises “1x50 Conv2d,16” module 301, “BatchNorm” module 302, “LeakyRule,0.2” module 303, “62x1 Conv2D, 32” module 304, “Flatten” module 305, “Dense,32” module 306, “Dense,16” module 307, “Reparameterize, (μ,σ) ” module 308, “62x1 ConvTranspose2D,32” module 309, “1x50 ConvTranspose2D,16” module 310 and “5x5 ConvTransposed2D,1” module 311, and the arrangement of these modules is illustrated in Figure 3. The detailed workings of each of these modules are omitted for brevity as they are known to one skilled in the art.

[0032] Trained VAE model 104 is able to capture key discriminating features relating to different classes of input data as model 104 utilizes temporal convolutions followed by spatial filters to process the data before dimensional reduction is performed on the data. Based on the above approach, it was found that trained VAE model 104 was able to achieve low

reconstruction errors whilst capturing the latent features of the input data therefore allowing for latent features of input data to be extracted without the need for labels.

[0033] In embodiments of the disclosure, the VAE model may be trained based on a plurality of unlabeled non-target data, e.g., having a sample size between 1 to infinity, to identify and compare the shift in latent domains between target data and non-target data. As for the target data, in embodiments of the disclosure, the target data may be split into offline/online sessions and further split into different phases based on different recording dates as illustrated in Figures 4a and 4b, whereby each phase may comprise of, but is not limited to, 100 trials. For effective utilization of data from various subjects and datasets, the datasets may be organized into their respective categories. As shown in Figures 4a and 4b, the highest-level category is the subject, followed by distinct recording sessions, and ultimately, various data recording phases within a single session. This organization is intended to ensure that when the data selection process occurs, the smallest selected unit would possess a consistent distribution that aligns with the specific domain of interest.

[0034] In embodiments of the disclosure, the method of selecting target data that is to be used may comprise one of the following three methods. In the first method, the target data can be extracted from each individual subject (IS) as a whole. In the second method, individual blocks comprising 100 trial data can be selected across different subjects including the additional session data from the target subject. This offers finer selection of the data used. In the third method, a selection of individual blocks of 100 trial data may be used with the exclusion of the additional session data (e.g., the offline data shown in Figure 4a). This approach is more representative of real-world scenarios where labelled host subject data (i.e., target data) is difficult to obtain. For the second and third methods, data from a total of 100 blocks of 100 trials may be chosen as the fine-tuning data that is to be used.

[0035] Once trained, the trained VAE model may then be used to obtain key latent features of unlabeled non-target data by maximising the log-likelihood $\log p_{\theta}(N|z_n)$ while the KL divergence is used for regularisation. The evidence lower bound (ELBO) is thus given by:

$$ELBO \geq EQ_{\theta}(z_n|N)[\log p(N|z_n)] - KL(q_{\theta}(z_n|N)||q(z_n)) \quad \dots (3)$$

where $EQ_{\Theta}(z_n|N)[\log \rho(N|z_n)]$ represents the reconstruction loss of each of the non-target data N and z_n is defined as the extracted latent features of unlabeled non-target data and $KL(q_{\Theta}(z_n|N)||q(z_n))$ represents the KL divergence between the distribution of the variational posterior $q_{\Theta}(z_n|N)$ and the prior distribution of the latent features $q(z_n)$.

[0036] As the latent features z_n obtained (using the trained VAE model) from unlabeled non-target data are highly representative of brain activity such as EEG signals, a comparison in the latent space is then conducted between the obtained latent features z_n against the features z_x obtained from known labelled data or target data. This is done by using the trained VAE model to compute the similarity or similarity scores between the non-target data and target data.

[0037] To achieve this, the trained VAE model, which has the same weights Θ , is then used to infer the reconstruction losses of each of the target data and this reconstruction loss is defined as: $-EQ_{\Theta}(z_x|N)[\log \rho(X|z_x)]$. The KL divergence is also used to compare between the learned expected distribution, that was trained based on N -number of non-target data against the actual distribution of target data $q(z_x)$, given by the KL divergence function $KL(q_{\Theta}(z_x|N)||p(z_x))$.

[0038] By doing so, target data having the highest overall loss can therefore be said to have the furthest similarity in latent features, i.e., lower similarity scores, when compared to the latest non-target data. Instead of directly comparing the embeddings, the trained VAE module is used to obtain a distance metric since having close embeddings does not necessarily equate to close true distributions. Given the ill-defined and abstract nature of non-stationary domains associated with EEG signals, no explicit true label exists for the outcome of contrastive measurements thus, the contrastive distance computed is not backpropagated as a loss to maintain the unsupervised learning. Hence, the distance function of the non-target data to the target data may be given by the contrastive difference between the computed Gaussian distribution of the non-target data N and the computed Gaussian distribution of the target data X . The contrastive distance may be summarized as the sum of the reconstruction loss and the KL losses of the unlabeled non-target data N having the parameters q_{Θ} as learned from target data X and is defined as:

$$\text{Contrastive distance} = -EQ_{\Theta}(z_x|N)[\log \rho(X|z_x)] + KL(q_{\Theta}(z_x|N)||p(z_x)) \quad \dots(4)$$

[0039] It should be noted that the contrastive comparison described above was used towards determining adaptation compatibility of EEG data, i.e., non-target data, in order to build a robust adaptation-set. Additionally, the latent features of the EEG signals are compared directly through the computed overall loss of the network as opposed to a technique that directly compares the learnt embeddings.

[0040] Once the contrastive distances have been computed, target data having similar distributions as the learned latent distributions of the non-target data N would have smaller contrastive distances and therefore have amongst the lowest overall computed loss. Since the training of the VAE model optimises for the Gaussian probability distribution of the non-target dataset, it is possible to infer that target data having lower losses would have a closer true Gaussian probability distribution to the non-target data that the inference model was trained on given that $\log p_{\Theta}(N|z_n)$ is close to $\log p_{\Theta}(X|z_x)$ when the VAE model has weights Θ .

[0041] It should be noted that the trained VAE model should be applied on larger groups of data rather than individual data points, as the result obtained is a better measure of overall data and is less sensitive to noise resulting from changes in dimensionality. Therefore, the non-target data can be ranked whereby the non-target data with closest similarity in underlying domains to the target data has the lowest computed overall loss while the non-target data with furthest domain shift has the highest loss. Unlabeled non-target data having a contrastive distance above a defined threshold is thus discarded and the remaining unlabeled non-target data is used as the new training dataset to fine tune the trained VAE model.

Experiment

[0042] Before the trained VAE model is fine-tuned, a finer selection process is performed on the new training dataset before it is used. This is implemented by splitting each subject's data in the new training dataset into individual phases. The rationale for this is to take into consideration setting-to-setting variations as intra-subject variation is an important factor in EEG signal recognition. In embodiments of this disclosure, a phase refers to a block of 100 trials as referred to in Figures 4a and 4b whereby the all the 100 trials were collected in one

sitting. During the finer selection process, data from the new training dataset will be selected in blocks of 100 trials as opposed to the whole subject data being selected.

[0043] After employing the more selective data selection method described earlier, the next step involves fine-tuning the trained VAE model using the chosen data. In embodiments of this disclosure, we vary the number of non-target data to investigate the impact of different amounts of non-target data on the data selection process and subsequent adaptation results. One-shot adaptation is conducted when a sole unlabeled non-target datapoint is employed, while few-shot adaptation involves using a limited number of unlabeled non-target datapoint. In this experiment, the various adaptation performances were examined for up to 10 unseen and unlabeled non-target datapoints.

[0044] During the adaptation process, freezing of layers may additionally improve the time taken and efficiency of the adaptation. The number of computations for back propagation can be significantly reduced and as a result reduce the total time taken for the adaptation to complete. However, freezing too many layers may also result in a less than optimal increase in the performance of the adapted network. It was previously noted that freezing only certain layers of the network leads to the best overall improvement in network performance during adaptation. Similarly, this experiment seeks to examine the ideal freezing scheme to be used in conjunction with the model re-learning framework.

[0045] Scheme 1 of this experiment freezes most of the classification network's layers, while Scheme 5 freezes none of the network layers. In this experiment, the effects of freezing different layers will be analyzed and the effects of using a different number of non-target data for the proposed adaptation framework will be analyzed as well.

Neural network models

[0046] The VAE architecture illustrated in Figure 3 is used for modeling the semantic domains of the target EEG signals and for comparing non-target data and target data for data selection. It should be noted that the network architecture is shared between both stages. The baseline classifier used in this experiment follows the DeepConvNet architecture which is used for subsequent adaptation and the models or neural networks were trained using a NVIDIA

GeForce RTX 2080 Ti Graphics Processing Unit, with an Intel i7-10700 Core Processing Unit (2.90GHz).

Dataset

[0047] To evaluate the effectiveness of the proposed adaptive framework, the EEG dataset from the Department of Brain and Cognitive Engineering, Korea University was utilized. The utilized dataset comprised of 54 healthy subjects (ages 24–35) performing binary class MI tasks. The EEG signals were recorded using BrainAmp (Brain Products; Munich, Germany) with 62 Ag/AgCl electrodes at a sampling rate of 1000 Hz. Throughout the sampling process, all subjects remained anonymous, and the design of the experiments followed the well-established protocol set by Pfurtscheller and Neuper. Further details about the data and protocols can be found in Lee, M.-H., Kwon, O.-Y., Kim, Y.-J., Kim, H.-K., Lee, Y.-E., Williamson, J., Fazli, S., and Lee, S.-W. “EEG dataset and OpenBMI toolbox for three BCI paradigms: An Investigation into BCI Illiteracy”. GigaScience, 8(5):Giz002, 2019. Each subject participated in two data recording sessions across two different days with a total of 400 trials. The first session consisted of offline training for recording data and training a baseline classifier, and the second session involved online testing where visual feedback is provided to the subject by decoding live data via the classifier. Each session consisted of 200 trials, split into two phases of 100 trials. Pre-processing of the data was done by decimating the data to 1000 data points per trial.

Metrics

[0048] An identical number of test trials were used for benchmarking state of the art models against the adaptive framework described in this disclosure. Baseline models are evaluated on a subject-independent basis whereby the test set consists of the last 100 trials of the online phase. Similarly, adapted models are evaluated based on the last 100 trials less the trials used for few-shot learning. The accuracy of both baseline and adapted models are given by the total number of correct classifications over the number of evaluation trials.

Results

Hyperparameter Analysis

[0049] To examine the performance of the adaptive framework, the various computational layers are frozen progressively while performing adaptation. Besides determining the number of layers that should be frozen, the number of trials to be used for model adaptation has to be determined as well. For fair comparison transfer learning for 1 to 10 trials was implemented using the source code disclosed in Zhang, K., Robinson, N., Lee, S.-W., and Guan, C. Adaptive Transfer Learning for EEG Motor Imagery Classification with Deep Convolutional Neural Network. *Neural Networks*, 136:1–10, 2021 (“**Zhang et. al. 2021**”). The scheme progressively freezes the layers of the network with Scheme 5 not freezing any layers and Scheme 1 freezing all layers except the for final fully connected layer. Freezing of the network layers gives the advantage of requiring fewer computations during the adaptation process, thus allowing for faster overall re-calibration times.

Adaptation Trials	Freezing Adaptation Scheme					
	Base	1	2	3	4	5
1	84.47	84.64	85.39	86.00	86.61	88.23
2	84.35	84.56	84.81	86.04	86.59	87.62
3	84.31	84.71	84.99	86.37	86.22	87.30
4	84.34	84.41	85.20	85.96	86.74	87.06
5	84.44	84.80	85.13	85.42	86.61	88.05
6	84.42	84.77	85.40	85.62	86.44	87.69
7	84.27	84.71	84.99	85.13	86.70	87.46
8	84.20	84.24	84.60	85.21	86.59	87.72
9	84.17	84.25	84.92	85.65	86.27	87.79
10	84.26	84.20	85.02	85.39	86.35	87.35

TABLE 1

[0050] From the results shown in Table 1, it was found that when Scheme 5 was used alongside the proposed adaptive framework, this resulted in the best overall improvement in the performance of the model. Furthermore, it is noticed that as the number of frozen layers decreases, the overall accuracy across the trials displayed an upward trend (Table 1). This differs from the previous state-of-the-art results which found that Scheme 4 gave an overall more positive result. We postulate that this occurred due to the notable increase in adaptation data by the proposed framework. It is intuitive that updating more weights of the entire network would result in overall faster convergence to the adaptation data. This is further supported by

the observation that across the different network freezing schemes, there is still a noteworthy upward improvement trend found when the proposed adaptive framework is used in Schemes 2-5. Thus, for the remainder of the experiment, the analysis is only performed on Scheme 5 (in which no layers were frozen).

[0051] From the experimental data, it was found that when a total of 1 target subject trial was used alongside Scheme 5 (Table 1), the best classification accuracy (88.23%) was achieved. Across the different number of few-shot trials, the proposed unsupervised framework consistently resulted in a significant improvement against the baseline ($p < 0.05$). This strongly suggests that although only one trial was used, the VAE model can sufficiently capture generalizable latent feature representations of the current target subject EEG signal paradigm that can subsequently be used to perform data selection for model re-learning.

[0052] In general, it was found that the method proposed in this disclosure offered improved performance over the original adaptation method with the best results achieving 88.23% and 86.89% respectively (Table 2 and 4). Overall, the difference in performance between the methods indicates that the method proposed in this disclosure is superior in identifying suitable data for fine-tuning rather than relying on the assumption that the adaptation data is representative of the target data.

Adaptation Trials	Data Selection Scheme				
	Base	SoTA	IS	IP (Include)	IP (Exclude)
1	84.47	85.90	85.48	85.19	88.23
2	84.35	85.77	86.23	85.68	87.62
3	84.31	85.64	85.82	85.34	87.30
4	84.34	85.73	85.92	85.30	87.06
5	84.44	85.79	85.89	85.73	88.05
6	84.42	85.76	86.27	85.82	87.69
7	84.27	85.66	85.67	84.97	87.46
8	84.20	85.65	85.79	85.06	87.72
9	84.17	85.61	86.11	85.41	87.79
10	84.14	83.31	85.41	84.94	87.35

TABLE 2

[0053] Table 2 above sets out the accuracy results (%) of different data selection methods whereby Scheme 5 was used for the computation. In the table, “SoTA” is defined as State-of-

the-art, “IS” is defined as Individual Subject, “IP” is defined as Individual Phases including or excluding target’s data.

[0054] When the possible methods of creating the adaptation dataset were compared (as shown in Table 2), it was found that when the individual phases of 100 trial blocks were selected, this resulted in the best overall improvement in the performance of the adaptive model. When whole subjects and individual trials were selected, the results obtained were not as good. Since there are strong intra- and intersubject variability which may cause poorer model learning when used for fine-tuning, the proposed framework seeks to reduce domain gaps between adaptation and target data. Whole subjects might result in too much undesirable trials being selected alongside the desired trials, whereas individual trials may be overfit towards the target. Therefore, it was found that a phase block of 100 trials offers the best balance between closeness in domain features while still offering good generalizable training data. Interestingly, it is noted that the model performs better when the supplementary data was excluded. This strongly suggests that the target subjects often contain high inter-session variability in the signals, leading to a decrease in the model’s performance whenever the target subject additional data is chosen. Thus, it might be more beneficial to utilize data from other subjects which offer better generalizability and are more readily available.

Comparison of Pre-adaptation Baselines

[0055] Table 3 sets out a comparison in the Mean and Standard Deviation of the baseline models for a DeepConvNet model for subject independent modelling against the adaptive model described in this disclosure.

	Original (Zhang et. al., 2021)	Proposed adaptive model
Mean (SD)	84.19 (± 9.98)	84.47 (± 11.97)

TABLE 3

[0056] It is noted that the accuracy given by the baseline model using the proposed adaptive model (84.47%) offered a close similarity in accuracy against the state-of-the-art model’s best baseline (84.19%) (Zhang et al., 2021). Currently, the training of the baseline subject independent model offers no set procedure to determine the training and validation split of the

known data. The baseline models are found using n-fold cross validations and using the best model on the validation set for evaluation. However, a disadvantage of this is that there is no way to determine the best number of folds. This may result in variations of model performance across the different subject-splits dependent on the permutations due to non-trivial variations between subjects. Additionally, the amount of training time needed to form the baseline models would increase significantly with the number of folds.

[0057] Hence, the proposed subject-independent selective framework offers a robust and methodological way to train a strong and consistent baseline model in just a single iteration. This contrasts against performing n-fold random data splits which causes considerable escalation in model training times and may result in large variations in performance dependent on how the subjects are split.

Comparison of Adaptation Methods

[0058] Figures 5a and 5b illustrates a comparison of the proposed framework to the state-of-the-art across individual subjects' performance changes following an adaptation process where Figure 5a illustrates the individual subjects' performance for a subject-adaptive transfer learning framework (Zhang et. al., 2021) and Figure 5b illustrates the individual subjects' performance for the adaptive framework proposed in this disclosure. It can be seen from Figures 5a and 5b that when the improvement rates across all the individual subjects in the proposed method are examined, it is found that the number of subjects who had accuracy reductions after the adaptation process reduced from 12 to 8 subjects, whereby the only subjects that had drops in performance only suffered mostly from minor reductions in performance. Additionally, a greater number were able to benefit from a significantly higher increase in classification performance. This is in stark contrast compared to the current state-of-the-art whereby a wider proportion of subjects faced performance degradation following adaptation of the model with greater drops in accuracy and the improvements of the subjects were minimal.

Methodology	Mean (SD)	Median	Range (Max-Min)
CNN (Kwon et al., 2019)	74.15 (± 15.83)	75.00	60.00 (100.00-40.00)
EEG-GAN (Hartmann et al., 2018)	81.03 (± 9.97)	81.85	41.73 (98.77-57.04)
DCGAN (Fahimi et al., 2020)	81.85 (± 9.85)	81.40	39.84 (98.52-58.68)

Cycle-GAN (Yin et al., 2022)	82.34 (± 8.98)	81.52	34.05 (99.08-65.03)
Deep CNN (Schirrmester et al., 2017)	84.19 (± 9.98)	84.50	47.50 (99.50-52.00)
Unsupervised Few-shot Adaptation Baseline (Ours)	84.47 (± 11.97)	86.32	42.11 (100-57.89)
Unsupervised Few-shot Adaptation + 160 Target Trials (Ours)	85.82 (± 11.05)	89.36	39.36 (100-60.64)
Subject-Adaptive Transfer Learning + 160 Target Trials (Zhang et al., 2021)	86.89 (± 11.41)	88.50	44.00 (100-56.00)
Unsupervised Few-shot Adaptation + 1 Unlabelled Target Trial (Ours)	88.23 (± 10.36)	90.91	40.40 (100-59.60)

TABLE 4

[0059] Table 4 compares between average classification accuracy (%) and standard deviation of the proposed adaptive framework with various other state of the art methods. Overall, as it can be seen from Table 4, the proposed method achieved significantly higher ($p < 0.05$) accuracy rates while providing much more consistent improvement in performance following adaptation compared to the current state-of-the-art methods by offering an improvement in accuracy regardless of the scheme being used while utilizing fewer labels from the target. Other adaptive and augmentation methods that were based on the generative adversarial network (GAN) architecture offered only marginal improvements from the baseline. It is hypothesized that this might be because GAN-based networks do not take into consideration the latent features space and therefore struggle to discriminate data with ill-defined features that arise from subject-to-subject or session-to-session variations against useful data. The resultant model would therefore still be trained or fine-tuned using undesirable data. Similarly, while subject-adaptive transfer learning (Zhang et al., 2021) seeks to reduce subject-to-subject variations, there still exist significant signal variations between sessions (as illustrated in Figures 6a and 6b) which results in less-than-ideal adaptive performance. In particular, Figures 6a and 6b illustrate a t-Distributed Stochastic Neighbour Embedding (t-SNE) plots of learned representations from a subject. In these plots, highly separable clusters show the presence of distinct signal variations between offline and online sessions which affects the efficacy of subject-adaptive transfer learning.

[0060] Therefore, it can be said that the proposed adaptive framework described in this disclosure has a significant improvement over the previous state-of-the-art in decoding performance, while simultaneously having reduced reliance on target labelled data.

Qualitative Analysis

[0061] When the t-SNE representations (as illustrated in Figures 6a and 6b) as extracted from the trained proposed VAE model in the framework are examined, it may be qualitatively ascertained as to why certain subjects observe a decrease in performance even when using within-subject data from earlier recordings are utilized to perform the adaptation process.

[0062] When the VAE model is applied across the entire subject's data across the four phases, the VAE model will attempt to learn common latent features. Upon examination of the learned representations of a single subject, it can be observed that there exists a non-trivial intra-subject variability in terms of the latent features across the four phases especially between offline and online sessions. The offline adaptation data used for transfer learning therefore does not belong to the same continuous distribution as the online evaluation data. This would strongly indicate that for these subjects, using earlier data from a separate session for adaptation may result in decrease in model performance. Overall, the high classification accuracy achieved by the baseline subject-independent model strongly indicates that there exist common stationary features that are generalizable across all the subjects. However, due to the non-stationary features associated with the varying signal collection settings, the adaptation of the model may be unable to identify the meaningful features from the additional data. Thus, for the subject-independent case, since large amounts of data from other subjects are used, the trained model learns to recognise meaningful latent representations that are common across the subjects that can be used to differentiate the signals.

[0063] To examine the variation of the valuation and evaluation trials against the fine-tuning dataset of the first 200 trials, the Savitzsky-Golay moving average of the mean Euclidean distance between the trial feature vectors were computed and plotted against the first 200 trial vectors from the first session. The average Euclidean distance between each target feature vector F_t and the set of 200 session 1 feature vectors $S_1, S_2, \dots, S_{200}$ may be expressed in mathematical notation as follows:

$$\bar{d}(F_t, S_1, S_2, \dots, S_{200}) = \frac{1}{200} \sum_{i=1}^{200} ||F_t - S_i||_2 \quad \dots (5)$$

[0064] Subsequently, a Savitzsky-Golay filter is applied on the computed points to obtain a smoother plot that displays the distance variation of each validation and evaluation trial from the fine-tuning set of the first session. Each new point is computed using the following formula:

$$\hat{d}_i = \sum_{j=-m}^m c_j S_{i+j} \quad \dots (6)$$

where m is the half-window size, and c_j are the filter coefficients. The filter coefficients are calculated using the least-squares method to fit a polynomial of degree n to the data within the window. The detailed workings of the Savitzsky-Golay filter are omitted for brevity as it is well known to one skilled in the art.

[0065] Contrary to intuition, the use of same-subject data for finetuning does not necessarily indicate an improvement in classification performance and may instead be detrimental. Figure 7 illustrates feature values plotted against time of learned representations obtained from a subject. In a continuous online setting, adaptation is done using sequential data while in a typical transfer learning scenario (Zhang et al., 2021), offline data is used for fine-tuning while the first half of online data is used for validation. Figure 8 illustrates a Savitzsky-Golay moving average of the Euclidean distance between the validation and evaluation trials against the trials in the transfer set where a higher distance indicates further deviation from the transfer set.

[0066] From the extracted features learnt from the network (as illustrated in Figures 7 and 8), it can be observed that the variation in the subject's learnt signal features display a strong correlation with the adaptation performance of the subject. This may be explained by the distinct variation between the training, validation, and evaluation phases. The difference between the transfer-learning set from the evaluation set indicates that the fine-tuned model would only be able to learn a distribution of latent features which are highly different to the evaluation set, causing a large drop in final decoding performance. In addition, during the fine-tuning stage, the large variation of the features from the validation trials to the evaluation set causes the model to learn weights that generalize to a distribution that does not follow the query dataset, resulting in a decrease in decoding accuracy. The measured variability within subjects using a contrastive distance metric is therefore an important aspect towards robust adaptive performance and is heavily under-utilized in present methods.

[0067] A process for fine tuning a trained neural network in accordance with embodiments of the present disclosure is illustrated in Figure 9 whereby process 900 may be carried out by computer system 100.

[0068] Process 900 begins at step 902 with process 900 predicting, using the trained neural network, latent features from a first plurality of unlabeled non-target data, whereby the neural network was trained based on a second plurality of unlabeled non-target data. At step 903, latent features from a plurality of non-target data are predicted by process 900. Process 900 then computes similarity values between the predicted latent features of the first plurality of unlabeled non-target data and the predicted latent features of the plurality of labeled target data at step 904. Process 900 then selects unlabeled non-target data that have similarity values below a defined threshold from the first set of unlabeled non-target data. This takes place at step 906. Finally, at step 908, process 900 then fine tunes the trained neural network by re-training the trained neural network using the selected unlabeled non-target data.

[0069] In embodiments of the disclosure, the neural network may comprise a convolutional neural network-variational autoencoder (CNN-VAE).

[0070] In embodiments of the disclosure, during the predicting of the latent features from a first plurality of unlabeled non-target data, process 900 may perform the prediction of the latent features using a few-shot contrastive learning technique.

[0071] In embodiments of the disclosure the neural network may be trained by maximizing a log-likelihood $\log_{\rho_{\Theta}}(N|z_n)$ and by regularizing the training using a Kullback-Leibler (KL) divergence, where ρ_{Θ} is defined as a predicted bias of the neural network, N is defined as the second plurality of unlabeled non-target data and z_n is defined as latent features of the second plurality of unlabeled non-target data.

[0072] In embodiments of the disclosure, during the process of computing the similarity values between the predicted latent features of the first plurality of unlabeled non-target data and latent features of a plurality of labeled target data, process 900 may compute contrastive distances between a computed Gaussian distribution of the first plurality of unlabeled non-target data and a computed Gaussian distribution of the plurality of labeled target data, wherein the similarity values are defined as the computed contrastive distances.

[0073] In embodiments of the disclosure, the contrastive distances between the computed Gaussian distribution of the first plurality of unlabeled non-target data and the computed Gaussian distribution of the plurality of labeled target data may defined as:

$$\text{Contrastive distance} = -EQ_{\theta}(z_x|N)[\log_{\rho}(X|z_x)] + KL(q_{\theta}(z_x|N)||q(z_x))$$

where $-EQ_{\theta}(z_x|N)[\log_{\rho}(X|z_x)]$ is defined as a reconstruction loss function and $KL(q_{\theta}(z_x|N)||q(z_x))$ is defined as a Kullback-Leibler (KL) divergence function.

[0074] Numerous other changes, substitutions, variations, and modifications may be ascertained by the skilled in the art and it is intended that the present application encompass all such changes, substitutions, variations, and modifications as falling within the scope of the appended claims.

CLAIMS:

1. A computer system configured to fine-tune a trained neural network, the computer system comprising:
a computer processor; and
a non-transitory media readable by the processing unit, the media storing instructions that when executed by the processing unit, causes the processing unit to:
 predict, using the trained neural network, latent features from a first plurality of unlabeled non-target data, whereby the neural network was trained based on a second plurality of unlabeled non-target data;
 predict, using the trained neural network, latent features of a plurality of labeled target data;
 compute similarity values between the predicted latent features of the first plurality of unlabeled non-target data and the predicted latent features of the plurality of labeled target data;
 select unlabeled non-target data that have similarity values below a defined threshold from the first set of unlabeled non-target data; and
 fine-tune the trained neural network by re-training the trained neural network using the selected unlabeled non-target data.
2. The computer system according to claim 1, whereby the neural network comprises a convolutional neural network-variational autoencoder (CNN-VAE).
3. The computer system according to claim 1, whereby the instructions to predict latent features from a first plurality of unlabeled non-target data comprises instructions for directing the processing unit to:
 perform the prediction of the latent features using a few-shot contrastive learning technique.
4. The computer system according to claim 1, whereby the neural network was trained by maximizing a log-likelihood $\log_{\rho_{\theta}}(N|z_n)$ and by regularizing the training using a Kullback-Leibler (KL) divergence, where ρ_{θ} is defined as a predicted bias of the neural

network, N is defined as the second plurality of unlabeled non-target data and z_n is defined as latent features of the second plurality of unlabeled non-target data.

5. The computer system according to any one of claims 1 to 4, whereby the instructions to compute the similarity values between the predicted latent features of the first plurality of unlabeled non-target data and latent features of a plurality of labeled target data comprises instructions for directing the processing unit to:

compute contrastive distances between a computed Gaussian distribution of the first plurality of unlabeled non-target data and a computed Gaussian distribution of the plurality of labeled target data, wherein the similarity values are defined as the computed contrastive distances.

6. The computer system according to claim 5, wherein the contrastive distances between the computed Gaussian distribution of the first plurality of unlabeled non-target data and the computed Gaussian distribution of the plurality of labeled target data is defined as:

$$\text{Contrastive distance} = -EQ_{\Theta}(z_x|N)[\log_{\rho}(X|z_x)] + KL(q_{\Theta}(z_x|N)||q(z_x))$$

where $-EQ_{\Theta}(z_x|N)[\log_{\rho}(X|z_x)]$ is defined as a reconstruction loss function and $KL(q_{\Theta}(z_x|N)||q(z_x))$ is defined as a Kullback-Leibler (KL) divergence function.

7. A method to fine-tune a trained neural network using a computer system, the method comprising:

predicting, using the trained neural network, latent features from a first plurality of unlabeled non-target data, whereby the neural network was trained based on a second plurality of unlabeled non-target data;

predicting, using the trained neural network, latent features of a plurality of labeled target data;

computing, using the computer system, similarity values between the predicted latent features of the first plurality of unlabeled non-target data and the predicted latent features of the plurality of labeled target data;

selecting, using the computer system, unlabeled non-target data that have similarity values below a defined threshold from the first set of unlabeled non-target data; and

fine-tuning, using the computer system, the trained neural network by re-training the trained neural network using the selected unlabeled non-target data.

8. The method according to claim 7, whereby the neural network comprises a convolutional neural network-variational autoencoder (CNN-VAE).
9. The method according to claim 7, whereby the predicting of the latent features from a first plurality of unlabeled non-target data comprises the steps of:
 - performing the prediction of the latent features using a few-shot contrastive learning technique.
10. The method according to claim 7, whereby the neural network was trained by maximizing a log-likelihood $\log_{\rho_{\theta}}(N|z_n)$ and by regularizing the training using a Kullback-Leibler (KL) divergence, where ρ_{θ} is defined as a predicted bias of the neural network, N is defined as the second plurality of unlabeled non-target data and z_n is defined as latent features of the second plurality of unlabeled non-target data.
11. The method according to any one of claims 7 to 9, whereby the computing the similarity values between the predicted latent features of the first plurality of unlabeled non-target data and latent features of a plurality of labeled target data comprises the steps of:
 - computing contrastive distances between a computed Gaussian distribution of the first plurality of unlabeled non-target data and a computed Gaussian distribution of the plurality of labeled target data, wherein the similarity values are defined as the computed contrastive distances.
12. The method according to claim 10, wherein the contrastive distances between the computed Gaussian distribution of the first plurality of unlabeled non-target data and the computed Gaussian distribution of the plurality of labeled target data is defined as:

$$\text{Contrastive distance} = -EQ_{\theta}(z_x|N)[\log_{\rho}(X|z_x)] + KL(q_{\theta}(z_x|N)||q(z_x))$$
 where $-EQ_{\theta}(z_x|N)[\log_{\rho}(X|z_x)]$ is defined as a reconstruction loss function and $KL(q_{\theta}(z_x|N)||q(z_x))$ is defined as a Kullback-Leibler (KL) divergence function.
13. A computer system configured to fine-tune a trained neural network, the computer system comprising:

a neural network module comprising the trained neural network, whereby the neural network was trained based on a second plurality of unlabeled non-target data, the neural network module being configured to:

predict latent features from a first plurality of unlabeled non-target data;

predict latent features of a plurality of labeled target data;

compute similarity values between the predicted latent features of the first plurality of unlabeled non-target data and the predicted latent features of the plurality of labeled target data;

select unlabeled non-target data that have similarity values below a defined threshold from the first set of unlabeled non-target data; and

fine-tune the trained neural network by re-training the trained neural network using the selected unlabeled non-target data.

14. The computer system according to claim 13, whereby the neural network comprises a convolutional neural network-variational autoencoder (CNN-VAE).

15. The computer system according to claim 13, whereby the predicting of the latent features from a first plurality of unlabeled non-target data comprises the neural network module being configured to:

perform the prediction of the latent features using a few-shot contrastive learning technique.

16. The computer system according to claim 13, whereby the neural network was trained by maximizing a log-likelihood $\log_{\rho_{\theta}}(N|z_n)$ and by regularizing the training using a Kullback-Leibler (KL) divergence, where ρ_{θ} is defined as a predicted bias of the neural network, N is defined as the second plurality of unlabeled non-target data and z_n is defined as latent features of the second plurality of unlabeled non-target data.

17. The computer system according to any one of claims 13 to 16, whereby the computing the similarity values between the predicted latent features of the first plurality of unlabeled non-target data and latent features of a plurality of labeled target data comprises the neural network module being configured to:

compute contrastive distances between a computed Gaussian distribution of the first plurality of unlabeled non-target data and a computed Gaussian distribution of the

plurality of labeled target data, wherein the similarity values are defined as the computed contrastive distances.

18. The computer system according to claim 17, wherein the contrastive distances between the computed Gaussian distribution of the first plurality of unlabeled non-target data and the computed Gaussian distribution of the plurality of labeled target data is defined as:

$$\text{Contrastive distance} = -EQ_{\theta}(z_x|N)[\log_{\rho}(X|z_x)] + KL(q_{\theta}(z_x|N)||q(z_x))$$

where $-EQ_{\theta}(z_x|N)[\log_{\rho}(X|z_x)]$ is defined as a reconstruction loss function and $KL(q_{\theta}(z_x|N)||q(z_x))$ is defined as a Kullback-Leibler (KL) divergence function.