Retain and Adapt: Online Sequential EEG Classification With Subject Shift

Tiehang Duan, Zhenyi Wang, Li Shen, Gianfranco Doretto, Donald A. Adjeroh, Fang Li, and Cui Tao

***Abstract*—Large variance exists in Electroencephalogram (EEG) signals with its pattern differing significantly across subjects. It is a challenging problem to perform online sequential decoding of EEG signals across different subjects, where a sequence of subjects arrive in temporal order and no signal data is jointly available beforehand. The challenges include the following two aspects: 1) the knowledge learned from previous subjects does not readily fit to future subjects, and fast adaptation is needed in the process; and 2) the EEG classifier could drastically erase information of learnt subjects as learning progresses, namely catastrophic forgetting. Most existing EEG decoding explorations use sizable data for pretraining purposes, and to the best of our knowledge we are the first to tackle this challenging online sequential decoding setting. In this work, we propose a unified bi-level meta-learning framework that enables the EEG decoder to simultaneously perform fast adaptation on future subjects and retain knowledge of previous subjects. In addition, we extend to the more general subject-agnostic scenario and propose a subject shift detection algorithm for situations that subject identity and the occurrence of subject shifts are unknown. We conducted experiments on three public EEG datasets for both subject-aware and subject-agnostic scenarios. The proposed method demonstrates its effectiveness in most of the ablation settings, e.g. an improvement of 5.73% for forgetting mitigation and 3.50% for forward adaptation on SEED dataset for subject agnostic scenarios.**

***Impact Statement*—Decoding of EEG signals is useful for translating brain signals directly into commands controlling devices such as robotic arms and wheelchairs, etc. Classic EEG decoding settings assume there is an ample amount of data available for pretraining purposes. This is not always available in real-world scenarios. Per our knowledge, this work is the first to tackle the challenging setting of online sequential EEG decoding with subject shifts, where subjects arrive sequentially and no data is jointly available beforehand. We propose a principled approach that simultaneously tackles challenges of fast adaptation and knowledge retaining during the process. The proposed model has potential wide applicability in real-world scenarios such as robotic arm navigation and rehabilitation training.**

***Index Terms*—Brain–computer interface (BCI), continual learning, EEG classification, transfer learning.**

I. Introduction

BRAIN–COMPUTER interfaces (BCI) analyze human brain activities through processing and decoding of recorded brain signals [1]. The general method of acquiring such signals is to use Electroencephalography (EEG) equipment, with the privilege of being noninvasive, having high resolution across the temporal axis, and relatively low cost to equip and maintain. Analysis of EEG signal provides useful information for translating it into control commands, and is currently widely used in clinical applications to help individuals with mild to severe motor disabilities including: 1) autonomous navigation of mechanical devices such as robotic arms and wheelchairs [2]; 2) control of digital interfaces such as mobile phone apps and sensors of smart homes [3]; and 3) help detect medical conditions related to irregular brain activity [4].

Classic cross-subject and intrasubject EEG decoding settings assume there is an ample amount of signal data for pretraining purposes. This is not always available in real-world decoding scenarios as current EEG datasets only consist of a limited number of classes (e.g., DEAP dataset has four different classes of actions), while in real-world applications such as digital tablet control or wheelchair control, the actions to be decoded by the BCI system is much more. It is nearly impossible to collect enough pretraining data for all the different actions. It is therefore important for the BCI system to conduct online decoding from a cold start. Additionally, previous explorations have shown the patterns of EEG signal are significantly different across different subjects [5]. This poses challenges to the model to make accurate predictions when subject shift occurs during the online decoding process.

We formulate the problem setting as *online sequential EEG classification*, where the EEG decoder needs to make accurate real-time predictions on different subjects arriving sequentially, with underlying data distribution constantly changing. An illustration of this scenario is shown in Fig. 1. This brought significant challenges to the EEG decoder in terms of two

Manuscript received 2 January 2024; revised 1 March 2024; accepted 29 March 2024. Date of publication 5 April 2024; date of current version 10 September 2024. This work was supported in part by the U.S. National Science Foundation under Grant 1920920, Grant 2125872, and Grant 2223793, in part by the American Heart Association (AHA) under Grant 19GPSGC35180031, and in part by the National Institutes of Health (NIH) under Grant R01AG084236 and Grant R01AG083039. This article was recommended for publication by Associate Editor Chin Teng Lin upon evaluation of the reviewers' comments. *(Corresponding authors: Donald A. Adjeroh; Cui Tao.)*

Tiehang Duan, Fang Li, and Cui Tao are with the Department of Artificial Intelligence and Informatics, Mayo Clinic, Jacksonville, FL 32216 USA (e-mail: tiehang.duan@gmail.com; li.fang@mayo.edu; tao.cui@mayo.edu).

Zhenyi Wang is with the University of Maryland, College Park, MD 20740 USA (e-mail: wangzhenyineu@gmail.com).

Li Shen is with JD Explore Academy, Beijing 100101, China (e-mail: mathshenli@gmail.com).

Gianfranco Doretto and Donald A. Adjeroh are with Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506 USA (e-mail: gianfranco.doretto@mail.wvu.edu; donald.adjeroh@mail.wvu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAI.2024.3385390, provided by the authors.

Digital Object Identifier 10.1109/TAI.2024.3385390

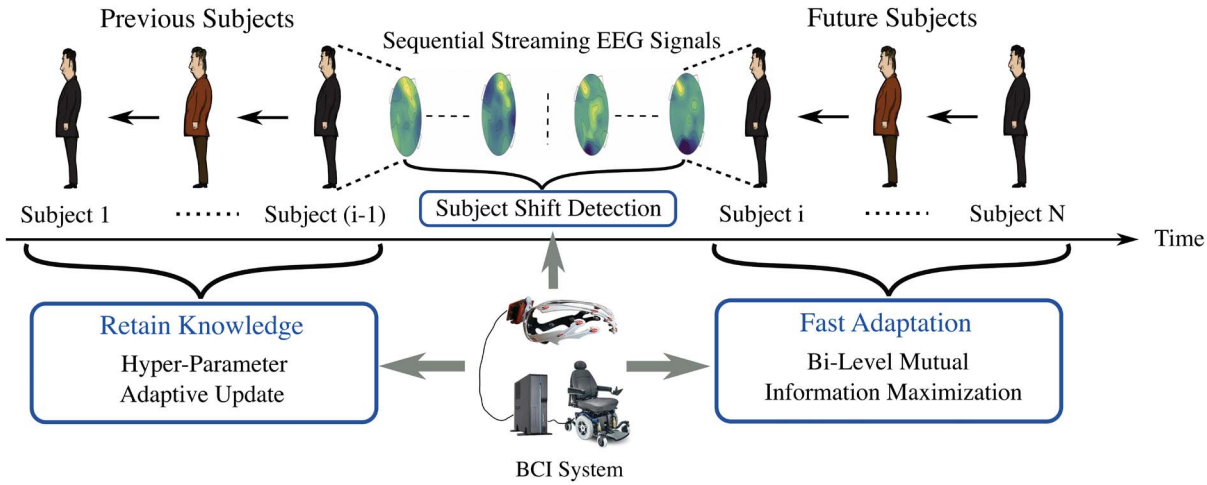

Fig. 1. Illustration of the online sequential EEG decoding scenario, with subjects arriving sequentially and produce streaming EEG data to the BCI system. The BCI system needs to simultaneously fast adapt onto future subjects and retain knowledge of previous subjects.

aspects: 1) with no data available for the model to jointly train beforehand, the EEG decoder needs to quickly adapt previously learned knowledge to newly arriving subject; 2) the decoder needs to mitigate the catastrophic forgetting issue and retain knowledge of previous subjects.

Direct utilization of previously proposed EEG decoding approaches are not feasible to this problem setting. Some earlier works adopt participant-dependent model tuning and perform calibration for each subject [6], [7]. A recent thread of research focuses on building subject-independent EEG decoding models [8], [9], [10]. This alleviates the need for fine-tuning of each subject at the cost of moderate performance downgrade [11]. Moreover, these models function on the premise of a fully constructed training dataset. In this work, we alleviate this requirement and the model decodes streaming EEG data from different subjects in an online manner. Per our knowledge, we are the first to simultaneously tackle fast adaptation and forgetting mitigation issues under this challenging setting.

Here, we propose a learning-to-learn approach using an *a*daptive *m*eta optimizer with *b*i-level *m*utual information maximization (AMBM) for tackling online sequential decoding of EEG signals. The meta optimizer is formed as a bi-level optimization process, with training performed in the unit of meta episodes. Each episode consists of iterations of inner base loop followed by the outer meta loop. In base loop, it maximizes mutual information between original signal and extracted feature for holistic feature extraction. In meta loop, mutual information maximization between current subject and previous subjects is performed for fast adaptation. The overall workflow of AMBM model is shown in Fig. 2. Hyperparameter is adaptively updated in meta loop to avoid forgetting previous knowledge. It preserves parameters that are important to decoding previous subjects from being overwritten, by utilizing the gradient information from the memory buffer and current streaming data to decide the level of catastrophic interference between current subject and previous subjects on model parameters. We kept a small memory buffer to store data samples from

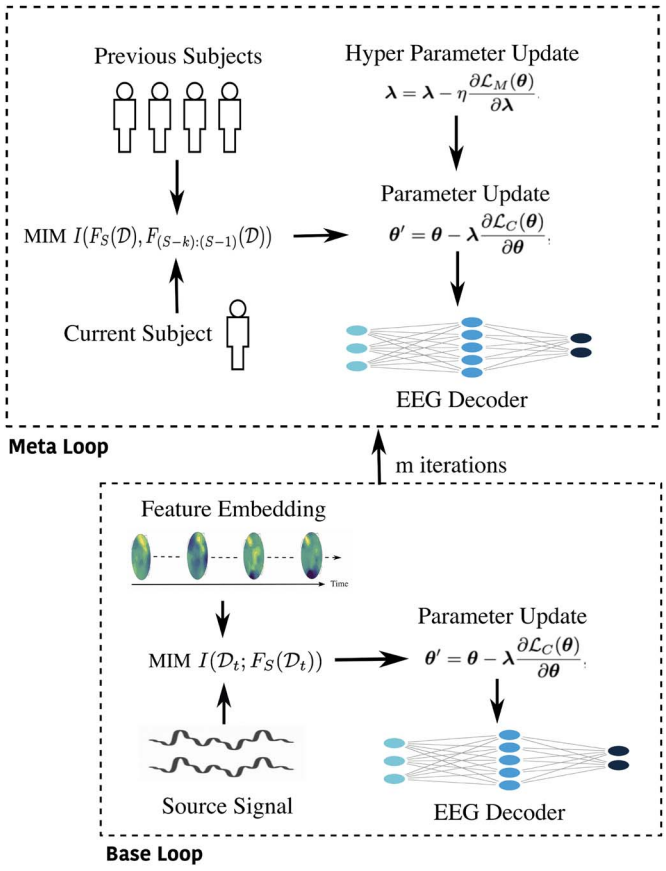

Fig. 2. Illustration on the workflow of AMBM model for simultaneous fast adaptation and knowledge retaining during online sequential EEG decoding.

previous subjects, which is used for both fast adaptation and forgetting mitigation.

For proper evaluation, we formulate benchmarks that mimic real-world BCI applications of subjects sequentially arriving in an online manner, with detailed ablation studies on different

sequence formations. The proposed method is first applied to the simplified sequential learning case that subject identity and shift boundary are known. We then extend the method to the more general subject-agnostic case, where subject identity is unknown in the learning process. We propose a simple and effective subject shift detection mechanism for this scenario. Performance comparison of the proposed method with current widely used baselines on the constructed online sequential decoding benchmarks shows the model is significantly more effective to sequentially decode EEG signal for a long period of time, both for subject-aware and subject-agnostic settings.

The contributions of this work are summarized as following.

1) We propose an adaptive meta optimizer with bi-level mutual information maximization to tackle the problem of online sequential EEG classification, applicable to a wide range of real-world scenarios.
2) We apply the proposed model to both subject-aware and subject-agnostic setting, depending on whether subject identity is known, and propose an efficient subject shift detection algorithm for the more general subject-agnostic setting.
3) We form proper benchmarks on top of public datasets for evaluation of the proposed method. We demonstrated the effectiveness of proposed approach with extensive empirical evaluation. The AMBM model achieved significant performance improvement compared to related strong baselines.

II. Related Work

A. EEG Classification

Signal processing methods are the classic way of decoding EEG signals [12]. A common approach is to go through bandpass filters and then perform feature extraction for decoding, e.g., [6] performed data preprocessing with filter bank (FB) and then extracted common spatial patterns for classification. Jatupaiboon et al. [13] extracted features based on frequency information from power spectral density (PSD), and performed classification with support vector machines (SVM). Covariance matrix is widely used in a large portion of the works as it provides relational information between the different EEG channels [14], [15], this can be useful for identifying patterns across the spatial distribution of brain signals. The performance of EEG decoding could be further improved by incorporating closed-loop neuro-modulation techniques [16], [17], which enables the subjects to perform adaptive reinforce or control of their brain activities.

Recently, researchers proposed neural EEG decoding approaches based on deep learning, which is formed with novel neural network structures for automatic feature extraction. EEGNet [18] is formed with three convolutional layers tailored for EEG decoding. It has low computational cost and is versatile across different BCI paradigms. CTCNN [19] is another state of the art EEG classifier formed with CNN which proposed a novel cropped training strategy. Zhang et al. [20] combines the cascade and parallel CNN structures by training multiple CNNs in parallel for improved performance. Explorations toward these directions largely focus on different types of model architectures and enable the decoding model to achieve state-of-the-art performance for classic decoding settings such as intrasubject decoding and cross-subject decoding, but less exploration is made to the more challenging online sequential decoding settings where no large-scale dataset is available for pretraining.

B. Continual Learning

Continual learning models [21], [22] aim to retain knowledge during learning from a sequence of different and dynamically evolving data distributions. The major challenge of this problem setting is known as catastrophic forgetting, where adaptation to a new data distribution deteriorates the model's performance on previous data distributions. The models in this field can be categorized into three categories: 1) Memory replay based methods [23], which stores a small amount of previous data in a memory buffer for joint consideration with current data. Aljundi et al. [24] proposed to replay samples that have the maximum interference with foreseen parameters update. Lopez-Paz and Ranzato [25] comes up with gradient episodic memory that effectively minimizes negative backward transfers. 2) Regularization based methods [26], which utilize regularization terms to preserve previously learned information. Kirkpatrick et al. [27] performs Bayesian estimation on the importance of model parameters and penalizes update of parameters that are important to previous tasks. Li and Hoiem [28] utilizes knowledge distillation as an additional regularization term and allows the model to work well on both old and new tasks. 3) Dynamic expandable networks [29], which expand the network with additional subnet architectures on the fly during the learning process (network size and memory cost are unknown at the beginning). Rusu et al. [30] proposes progressive network which instantiates a new column in the network for each new task. Yoon et al. [31] proposes to dynamically expand network capacity by splitting and duplicating existing network structures as learning progresses. The previous explorations for forgetting mitigation were mostly based on empirical study with the usage of memory buffer, expandable architecture and adoption of regularization terms. In this work we proposed a principled approach for forgetting mitigation by formulating it as an optimization problem on adaptive hyperparameters. Additionally, previous explorations mostly focus on either classic computer vision tasks (such as continual image deraining [32] and depth estimation [33], etc.) or natural language processing tasks (such as continual sentiment classification [34] and dialogue generation [35]). To our knowledge, little exploration has been made on its effectiveness to tackle subject shifts for clinical prediction tasks.

C. Meta Learning

Meta learning [36] offers promising solutions for a neural network to perform fast adaptation on an unseen task after learning across similar tasks. Specifically, [37] learns a metric space and fast adaptation is performed by computing distance toward prototype representations of each class. Finn et al. [38]

proposes a simple and effective approach for model and task agnostic meta-learning leveraging a double loop optimization structure. Rusu et al. [39] proposes to learn latent representation of model parameters conditioned on task data, which enables task-specific initialization of meta parameters for fast adaptation. Online meta learning [40] is proposed recently for better performance on tasks arriving in sequential order. Most previous works on meta learning is for nonevolving data distributions. Its effectiveness in sequential learning scenarios over multiple domains is currently under-explored. Additionally, its effectiveness for sequential subject adaptation in clinical prediction tasks remains largely unknown.

To the best of our knowledge, we are the first to tackle the challenging setting of online sequential EEG classification. As no dataset is available to pretrain the EEG decoder, traditional intrasubject and cross-subject EEG classifiers are not feasible to this challenging setting. Different from existing EEG decoding models with domain adaptation techniques, this work effectively tackles both challenges of fast adaptation and forgetting mitigation during online sequential EEG classification, utilizing a unified bi-level meta learning framework. Compared to classic continual learning approaches, the proposed method does not involve explicit memory replay and the hyperparameter adaptation functions in meta loop which only performs once for each meta episode, effectively reduces the computational complexity. Additionally, the proposed approach seamlessly integrates the bi-level mutual information maximization into the bi-level optimization loops and more effectively leverages the rich information at feature level, enabling faster adaptation compared to other meta learning approaches.

III. Methodology

A. Problem Setup

The definition of online sequential EEG classification is as follows:

Definition 1: Online Sequential EEG Classification. A sequence of subjects, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_J$, arrive sequentially and produce streaming EEG data to the BCI system. Each subject $\mathcal{S}_i$ produces a labeled sub dataset $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{B_i}$ with B_i labeled datapoints, where $\boldsymbol{x}_k$ is the signal segment and $\boldsymbol{y}_k$ the corresponding label. The subjects change over time which brings data distribution shift, and the EEG decoder should be able to *detect and adapt to this distribution change in real-time while preserving previously learned knowledge.*

We leverage the framework of model agnostic meta learning (MAML) for formulation of our approach. MAML is a general purpose gradient based meta learning approach for fast adaptation of tasks. It involves a bi-level structure, with the lower level or base loop targets on gradient accumulation of noisy data and higher level or meta loop for fast adaptation. As we described below, adopting this framework is beneficial for both fast adaptation to future subject and also preserving information of previous subjects.

We consider both subject-aware and subject-agnostic scenarios, with subject identity and subject shift boundary unknown for the latter case. The model is formed with a shared CNN-based network for feature extraction of all subjects, with the flexibility to expand a prediction head formed with a single layer for each arriving subject. Please note this network expansion operation shares similarities with previous methods such as [31], [41]. The purpose of keeping this prediction head is to: 1) create different views on subjects for bi-level mutual information maximization in forward adaptation, as detailed in Section III-B; and 2) avoid forgetting previous domain knowledge.

B. Fast Adaptation Toward Future Subjects

We perform bi-level mutual information maximization in the meta optimizer to learn features that are holistic and applicable to future subjects for online decoding.

For loss functions currently widely used for EEG decoding such as cross-entropy, it suffices for the model to learn discriminative features that properly separate different classes of current subject. There is no incentive for the model to learn features that are versatile and applicable to future subjects. Every time a new subject arrives, the model parameters need to be modified, resulting in slower adaptation. Here, we utilize mutual information for the model to learn holistic representations from the signal and increase versatility across subjects.

The process involves two phases. The first phase is to maximize the mutual information between signal data and extracted features. This is performed in base loop of meta optimization. The second phase maximizes the mutual information between different views of current and previous subjects. This happens in meta loop of the optimization.

The bi-level optimization can be formulated as

$$\mathrm{Alg}(\boldsymbol{\theta}, \mathcal{D}_t) = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \{\mathcal{L}_{ce}(\boldsymbol{\theta}, \mathcal{D}_t) - I(\mathcal{D}_t; F_s(\mathcal{D}_t))\} \quad (1)$$

$$\boldsymbol{\theta}^* := \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmin}} \left\{ \frac{1}{m} \sum_{t=T_{\mathcal{E}}}^{T_{\mathcal{E}}+m} \mathcal{L}(\mathrm{Alg}(\boldsymbol{\theta}, \mathcal{D}_t), \mathcal{D}) - I(F_s(\mathcal{D}), F_{(s-k):(s-1)}(\mathcal{D})) \right\}, \text{ where } \mathcal{D} = \{\mathcal{D}_t, \mathcal{D}_m\} \quad (2)$$

where $\mathcal{D}_t$ is streaming data of current subject and $\mathcal{D}_m$ is sampled from memory buffer (summary of variable notations is provided in Table I). m is the number of base loop iterations for each meta episode. $\mathcal{L}_{ce}(\cdot)$ is the cross entropy loss and $\mathrm{Alg}(\boldsymbol{\theta}, \mathcal{D}_t)$ is the updated parameter of base learner, which is initialized as meta parameters $\boldsymbol{\theta}^*$ of the last meta episode. $I(\mathcal{D}_t; F_s(\mathcal{D}_t))$ is the mutual information between signal segments and feature maps for holistic feature extraction, with $F_s(\cdot)$ being the network formed with prediction head of sth subject. $I(F_s(\mathcal{D}), F_{(s-k):(s-1)}(\mathcal{D}))$ is the mutual information aligning the different views of sequential subjects, for which we kept the most recent k prediction heads for adaptation purpose. We maintain a small memory buffer for data samples from previous subjects, with details as follows:

1) Memory Buffer Update: The model maintains a memory buffer (small in size) to keep samples of data from learnt subjects. The memory utilizes reservoir sampling, where each batch of EEG data has an equal probability of being selected.

For a memory buffer of size M, the first batches of EEG signal will be stored sequentially in the buffer until it is full.

TABLE I
SUMMARY OF VARIABLE NOTATIONS

Notation	Description
$\mathcal{D}_t$	Streaming data of current subject
$\mathcal{D}_m$	Sampled data from memory buffer
$\mathcal{L}_{ce}(\cdot)$	The cross entropy loss
$F_s(\cdot)$	Decoding network with prediction head of sth subject
$I(\cdot)$	Mutual information operator
$\boldsymbol{\theta}$	Parameters of base learner
$\boldsymbol{\theta}^*$	Parameters of meta learner
α	Learning rate of parameters
M	The size of memory buffer
$\Phi(\cdot)$	Batch normalization operation
B	Training batch size
$\boldsymbol{\lambda}$	Learning rate of hyperparameters
$\boldsymbol{e}_t$	Extracted feature map at time step t
$\boldsymbol{E}_t$	Moving average of feature maps at time step t
$\boldsymbol{d}_t$	Distance metric vector at time step t
$\mathcal{S}_t$	Denotation of the subject at time step t
l_{S_t}	Run length of subject $\mathcal{S}_t$
$U(.)$	Constant function depicting the prior of subject shift

For later batches of EEG data $\boldsymbol{x}^k$ arriving, a random number i is generated between 1 and k. If $i < M$, then $\boldsymbol{x}^k$ will be selected into memory and replace the data currently at ith place in memory. The memory kept data of size M from previous subjects in this way.

The memory buffer serves to: 1) extract versatile features across different subjects for fast adaptation; and 2) maintain performance of previous learned subjects with memory replay. More tailored design on memory data selection could potentially improve model performance, which we leave for future work.

2) Holistic Feature Extraction in Base Loop: The goal of this phase is for the feature extractor to extract useful information from source signal, namely *holistic* for its preserving of maximum information and reducing feature bias (with its measurement introduced in Appendix C). The maximization is performed between current signal segment $X \in \mathcal{D}_t$ and the features extracted $F(X) = \Phi(f_\theta(X))$, with $\Phi(\cdot)$ being the normalization operation.

It should be noted that directly computing this target is intractable, and following [42], we resort to maximizing its lower bound: mutual information on features extracted from different augmented signal data. Let X' be an augmented version of input data X created from data augmentation operations (details specified in Appendix B). Then we have

$$\begin{aligned}\max_\theta I(X; F(X)) &\geq \max_\theta I(F(X'); F(X)) \\ &\geq \log B + \text{InfoNCE}(g(\cdot); F(X); F(X'))\end{aligned} \tag{3}$$

where B is the batch size, InfoNCE is the contrastive loss function with

$$g(F(x_i), F(x'_j)) = \exp\left(\frac{F(x_i)^T F(x'_j)}{r}\right)$$

$$\text{InfoNCE}(g(\cdot); F(X); F(X')) = \frac{1}{B}\sum_{i=1}^{B} \log \frac{g(F(x_i), F(x'_i))}{\sum_{j=1}^{B} g(F(x_i), F(x'_j))}. \tag{4}$$

Algorithm 1 Online Sequential Adaptation

REQUIRE: Streaming EEG data from a sequence of subjects $\{\mathcal{D}_1, \ldots, \mathcal{D}_{N_1}; \ldots; \mathcal{D}_{N_i+1}, \ldots, \mathcal{D}_{N_{i+1}}; \ldots; \mathcal{D}_{N_{J-1}+1}, \ldots, \mathcal{D}_{N_J}\}$; where $\{N_i, i = 1, 2, \cdots, J-1\}$ are change of subject occasions; memory buffer $\mathcal{M}$
for $t = 1$ to N_J **do**
  sample $\mathcal{D}_m$ from $\mathcal{M}$, form training batch $\mathcal{D} = \{\mathcal{D}_t, \mathcal{D}_m\}$
  for iters of base loop **do**
    optimize $\boldsymbol{\theta}$ based on $\min_\theta\{\mathcal{L}_{ce}(\boldsymbol{\theta}, \mathcal{D}_t) - I(\mathcal{D}_t; F_s(\mathcal{D}_t))\}$
  end for
  optimize $\boldsymbol{\theta}^*$ based on $\min_{\theta^*}\{\frac{1}{m}\sum_{t=T_\mathcal{E}}^{T_\mathcal{E}+m} \mathcal{L}(\mathcal{A}lg(\boldsymbol{\theta}, \mathcal{D}_t), \mathcal{D}) - I(F_s(\mathcal{D}), F_{(s-k):(s-1)}(\mathcal{D}))\}$
  Reservoir sampling, $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{D}_t$ if sampled on $\mathcal{D}_t$
end for

$g(\cdot)$ can be seen as a similarity function with r a scaling factor. For proof of this lower bound please see Appendix H.

The equality to $\max_\theta I(F(X'); F(X))$ is achieved when $B \to \infty$ [43]. This suggests we can optimize model performance with increased batch size. However, for online streaming scenario of EEG decoding, it would cause delay in decoding as longer wait is needed to form a batch. A solution to this is to perform augmentation operations of the data in each batch, with more discussion provided in Appendix B.

3) Subject Adaptation in Meta Loop: After the model finished training in the base loop, the model is able to extract informative features of source signal. In meta loop, we want the model to learn to generalize and make predictions that are accurate across all subjects, both for adaptation and knowledge retaining purpose. The current streaming data $\mathcal{D}_t$ and samples from memory buffer $\mathcal{D}_m$ are jointly trained in meta loop. We add a lightweight prediction head formed with a single output perceptron layer for each subject, and the most recent k prediction heads are kept for adaptation purpose. We maximize $I(F_s(\mathcal{D}), F_{(s-k):(s-1)}(\mathcal{D}))$, with $I(F_s(\mathcal{D}), F_{(s-k):(s-1)}(\mathcal{D})) = \sum_{i=1}^{k} I(F_s(\mathcal{D}), F_{(s-i)}(\mathcal{D}))$ and $\mathcal{D} = \{\mathcal{D}_t, \mathcal{D}_m\}$. This helps the model to make accurate predictions on all subjects by aligning across the different subjects and improves the model's generalization ability. We provide the details of the adaptation mechanism in Algorithm 1.

C. Retain Knowledge of Previous Subjects

The proposed bi-level mutual information maximization allows the EEG decoder to fast adapt and achieve decent performance on current subject. However, such adaptation during online sequential learning results in erasing of model knowledge on previous subjects, also known as catastrophic forgetting. As illustrated in Fig. 3, the EEG decoder quickly forgets knowledge on subjects before A07 after sequential learning ends on all nine subjects. Here, we propose hyperparameter update in meta loop for the EEG decoder to dynamically adapt learning rate toward the parameters based on the learning progress of different portions of the network, which allows the model to preserve learned knowledge during online adaptation.

Classic way of optimizing neural networks for decoding streaming EEG data is through gradient descent

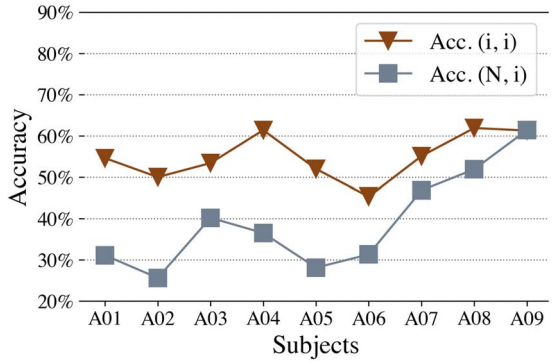

Fig. 3. Trial run of sequential EEG decoding on BCI IV-2a dataset reveals the catastrophic forgetting issue. Acc. (*i*, *i*) is the validation accuracy evaluated immediately after learning a subject, and Acc. (*N*, *i*) is the testing accuracy on *i*th subject after learning of all *N* subjects. The testing accuracy shows quick deterioration for subjects before A07 after finished learning of all nine subjects.

Algorithm 2 Online Sequential Knowledge Retain and Adaptation

REQUIRE: Streaming EEG data from a sequence of subjects $\{\mathcal{D}_1, \ldots, \mathcal{D}_{N_1}; \ldots; \mathcal{D}_{N_i+1}, \ldots, \mathcal{D}_{N_{i+1}}; \ldots; \mathcal{D}_{N_{J-1}+1}, \ldots, \mathcal{D}_{N_J}\}$; where $\{N_i, i = 1, 2, \cdots, J-1\}$ are change of subject occasions; Initialize learning rates $\boldsymbol{\lambda}_0$ and model parameters $\boldsymbol{\theta}$; η is the hyper parameter for adaptation on learning rate.
for $t = 1$ to N_J **do**
  sample $\mathcal{D}_m$ from $\mathcal{M}$, form training batch $\mathcal{D} = \{\mathcal{D}_t, \mathcal{D}_m\}$
  for iters of base loop **do**
    optimize $\boldsymbol{\theta}$ based on $\min_\theta\{\mathcal{L}_{ce}(\boldsymbol{\theta}, \mathcal{D}_t) - I(\mathcal{D}_t; F_S(\mathcal{D}_t))\}$
  end for
  optimize $\boldsymbol{\theta}^*$ based on $\min_{\theta^*}\{\frac{1}{m}\sum_{t=T_\mathcal{E}}^{T_\mathcal{E}+m} \mathcal{L}(\mathcal{A}lg\,(\boldsymbol{\theta}, \mathcal{D}_t)\,, \mathcal{D}) - I(F_s(\mathcal{D}), F_{(s-k):(s-1)}(\mathcal{D}))\}$
  $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \eta\frac{\partial \mathcal{L}_M(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\lambda}_t}$
  Reservoir sampling, $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{D}_t$ if sampled on $\mathcal{D}_t$
end for

$\boldsymbol{\theta} = \boldsymbol{\theta} - \boldsymbol{\lambda}(\partial\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_t)/\partial\boldsymbol{\theta})$ at step t, where $\boldsymbol{\theta}$ is the model parameter, $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_t)$ is the loss of current time step and $\boldsymbol{\lambda}$ is the learning rate. In standard setting, the learning rate is the same for all model parameters, however this would incur forgetting on portions of network that are important in decoding previous subjects. Here, we propose hyperparameter update in meta loop for the EEG decoder to dynamically adapt learning rate toward the parameters based on the learning progress of different portions of the network. Intuitively, parameters that are important to proper decoding of previous subjects will be adapted more gently and parameters with less influence will learn faster for efficient adaptation onto current subject. We utilize sampled data from memory buffer to measure the importance of parameters toward previous subjects. The interference of current data toward knowledge of previous subjects is then computed, with details below.

Defining $\nabla^t_{\boldsymbol{\theta}} = (\partial\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, \boldsymbol{y}_t)/\partial\boldsymbol{\theta})$ as the gradient at time step t. For two data points at time step i and j, the sign of dot product $\nabla^i_{\boldsymbol{\theta}} \cdot \nabla^j_{\boldsymbol{\theta}} = (\partial\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\partial\boldsymbol{\theta}) \cdot (\partial\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}_j, \boldsymbol{y}_j)/\partial\boldsymbol{\theta})$ indicates the potential interference between the two data points, e.g. $\nabla^i_{\boldsymbol{\theta}} \cdot \nabla^j_{\boldsymbol{\theta}} < 0$ suggests there is catastrophic interference between data of these two time steps, where learning on one will erase knowledge of the other; and $\nabla^i_{\boldsymbol{\theta}} \cdot \nabla^j_{\boldsymbol{\theta}} > 0$ suggests knowledge transfer and inheritage is possible. We utilize this information in design of the meta optimizer, where we perform the computation between the current streaming data and memory data, and make *learning rates learnable* for individual parameters based on this computation.

With $\mathcal{L}_M(\boldsymbol{\theta})$ the loss to optimize on generalization of memory data $\mathcal{D}_m$ and $\mathcal{L}_C(\boldsymbol{\theta})$ the loss to optimize on current streaming data $\mathcal{D}_t$, the update of learning rate is as follows:

$$\frac{\partial\mathcal{L}_M(\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}} = \frac{\partial\mathcal{L}_M(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\frac{\partial\boldsymbol{\theta}'}{\partial\boldsymbol{\lambda}} = -\frac{\partial\mathcal{L}_M(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} \cdot \frac{\partial\mathcal{L}_C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$$
$$\text{with } \boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{\lambda}\frac{\partial\mathcal{L}_C(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \tag{5}$$

$\boldsymbol{\theta}'$ is the updated parameter by one step on current loss and $\boldsymbol{\lambda}$ is the learning rate for adaptation. $\boldsymbol{\lambda}$ is updated following the derivation in (5) to be

$$\boldsymbol{\lambda} = \boldsymbol{\lambda} - \eta\frac{\partial\mathcal{L}_M(\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}}. \tag{6}$$

The RHS of (5) matches the intuitive explanation of catastrophic interference. With $\partial\mathcal{L}_M(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ being the gradient on memory data and $\partial\mathcal{L}_C(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ the gradient on current streaming data, their dot product being positive refers to the case that direction of $\partial\mathcal{L}_M(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ aligns with $\partial\mathcal{L}_C(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and offers knowledge transfer and inheritance. Otherwise, catastrophic interference happens and the learning rate is reduced for the parameter.

D. Subject Shift Detection for Subject Agnostic Setting

Here, we consider the more general setting that subject identity and corresponding subject shift are unknown, namely subject-agnostic setting. The major challenge in this setting is to detect the subject shift occurrences and properly expand the prediction heads on the model arch, which requires accurate subject shift detection with little latency. Here the challenges lie: 1) the volatile nature of EEG signals and their significant variance; and 2) different degrees of similarity across the subjects. Approaches such as directly setting up a threshold on the loss do not work based on our study. We adopt a Bayesian probabilistic approach in latent feature space to tackle this problem.

We denote the extracted feature map at time step t as $\boldsymbol{e}_t$. For streaming data of past p steps, moving average on the feature is $\boldsymbol{E}_t = \sum_{i=t-p}^{t}(\alpha_i \boldsymbol{e}_i)$ with $\sum_{i=t-p}^{t}(\alpha_i) = 1$, this reduces the variance in the signal. A distance metric vector is then formed based on most recent m steps of $\boldsymbol{E}_t$, denoted as $\boldsymbol{d}_t = (d(\boldsymbol{e}_t, \boldsymbol{E}_{t-1}), d(\boldsymbol{e}_t, \boldsymbol{E}_{t-2}), \cdots, d(\boldsymbol{e}_t, \boldsymbol{E}_{t-m}))$, with $d(\cdot)$ the distance metric. This encodes information for characterizing the subject across a wider range of time spans in a compressed manner.

We utilize the encoded vector $\boldsymbol{d}_t$ for subject shift detection. With S_t the current subject at time t, we want to get the posterior estimation on current run lengths l_{S_t}. $l_{S_t} = 0$ suggests subject

shift happens and $l_{S_t} = \tau > 0$ suggests the continued signal stream of current subject. The process is detailed as follows:

$$P(l_{S_t}|\boldsymbol{d}_{1:t}) \propto \sum_{l_{S_{t-1}}} P(l_{S_t}|l_{S_{t-1}})P(\boldsymbol{d}_t|l_{S_{t-1}}, \boldsymbol{d}_{1:t-1})P(l_{S_{t-1}}, \boldsymbol{d}_{1:t-1}). \quad (7)$$

With the first term being the prior of l_{S_t}, formed as

$$P(l_{S_t}|l_{S_{t-1}}) = \begin{cases} U(l_{S_{t-1}} + 1), & l_{S_t} = 0 \\ 1 - U(l_{S_{t-1}} + 1), & l_{S_t} = l_{S_{t-1}} + 1 \end{cases} \quad (8)$$

where $U(\cdot)$ is constant function. The upper term suggests subject shift and lower term denotes continuing on current subject. When subject shift happens, a prediction head is added onto the shared arch, and we perform the update $S_{t+1} = S_t + 1$.

IV. Empirical Evaluations

We evaluate the performance of proposed model by applying it to three different public EEG datasets, which are BCI IV-2a[1] [44], DEAP dataset[2] [45] and SEED dataset[3] [46], [47], for both subject-aware and subject-agnostic settings. The performance is evaluated in terms of two aspects: 1) fast adaptation to current subject; and 2) forgetting mitigation on previously learned subjects. We illustrate on construction of benchmark and metrics used for evaluation below.

A. Benchmark

We mimic the realistic BCI application scenario of subjects arriving in sequence. The subjects show varying degrees of similarity and difficulty, and we performed ablation study with three different subject arrival orderings: 1) sequence ordering based on id of subjects; 2) ascend ordering following the level of difficulty in decoding; and 3) descend ordering following the level of difficulty in decoding. The streaming EEG data is processed into batches of $t \times n$, with t the temporal span and n the number of channels. Details on data processing for each dataset are provided in Section IV-D.

B. Evaluation Metrics

For measuring the adaptation performance, we evaluate accuracy and AUC-ROC immediately after learning a new subject during online learning. We then evaluate the accuracy for all subjects after sequential learning ends, and use BWT to measure the forgetting mitigation performance. BWT is the abbreviation of backward transfer and measures the degree of catastrophic forgetting on all previous subjects at the end of sequential learning. $\text{BWT} = (1/N-1)\sum_{i=1}^{N-1} a_{N,i} - a_{i,i}$, with N being the total number of subjects in the sequence, and $a_{j,i}$ the accuracy on subject i after the model finished sequential learning on subject j. Negative value of BWT reveals catastrophic forgetting happens after learning the new subject, while positive value shows learning on new subject improves performance of previous subjects.

[1]http://bnci-horizon-2020.eu/database/data-sets
[2]https://www.eecs.qmul.ac.uk/mmv/datasets/deap/download.html
[3]http://bcmi.sjtu.edu.cn/seed/downloads.html

C. Baselines

The baselines that are incorporated into comparison are from three categories, including.

1) Lower and upper bound of model performance, with the lower bound named *classic sequential learning*, where base EEG decoder sequentially decodes the arriving subjects. The upper bound of model performance is *joint learning*, where data of all subjects are jointly available for the model to learn.

2) Domain adaptation techniques previously shown effective to cross subject EEG decoding, including the following: *MIDA-EEG* [48] creates a subspace with maximized independence between subjects. *TCA-EEG* [49] learns a reproducing Kernel Hilbert space (RKHS) across features of different subjects. *Deep-Transfer* [50] formed a deep CNN-LSTM network for transfer learning purpose. *RA-MDRM* [51] performs alignment on covariance matrix of subjects.

3) Currently widely used continual learning approaches, and combination of these approaches with proposed meta adaptation. The averaged gradient episodic memory (*A-Gem*) [26] constrain gradient directions to preserve learned knowledge. *HAL* [52] utilizes hindsight to mitigate forgetting on a series of previous anchor points. Experience replay (ER) [23] keeps samples of data in learned tasks to train together with current data. Gradient editing in memory (GMED) [53] makes stored samples hard to remember and mitigates overfitting. Maximally interfered retrieval (MIR) [24] replays examples with larger estimated interference.

D. Settings

1) Data Processing: BCI IV-2a dataset consists of nine subjects, with each subject went through a total of 576 trials. Each trial is processed into 22 (channels) by 400 (temporal span) segments. The adjacent segments have a stride size of 50 along the temporal axis. The data is categorized into four different classes following the four different types of motor imagery movements, namely left hand, right hand, tongue and both feet. We extracted the period of $t = 3$ to $t = 6$ s for each trial, within which the subject is actively performing motor imagery. This produces eight segments per trial at 250 Hz sampling rate.

For DEAP dataset, there are 32 subjects and each subject conducted 40 trials. Following the quadrants depicted by valence and arousal, the data is categorized into four different classes including: 1) high arousal and positive valence (HAPV); 2) low arousal and positive valence (LAPV); 3) high arousal and negative valence (HANV); and 4) low arousal and negative valence (LANV). We extracted the last 50 s of each trial in our experiment for improved data quality. Each trial is processed into 32 (channels) by 768 (temporal span) segments. The adjacent segments have a stride size of 128 and this produces 45 segments for each trial.

SEED dataset has a total of 15 subjects, with each subject participated in three sessions generating a total of 2775 samples. The recorded signals are labeled into three different classes, indicating the elicited positive, neural and negative emotions as subjects watching movie excerpts. The trials are processed into 62 (channels) by 800 (temporal span) segments, the size of

TABLE II
PERFORMANCE ON FORWARD ADAPTATION TOWARD SUBJECTS FOR DIFFERENT MODELS

Dataset	BCI IV-2a		DEAP		SEED	
Method	Accuracy	ROC-AUC	Accuracy	ROC-AUC	Accuracy	ROC-AUC
Sequential	48.12±0.24	68.57±0.15	42.56±0.49	61.91±0.37	52.28±0.39	68.34±0.25
Joint training	79.63±0.47	91.29±0.34	71.36±0.63	84.82±0.44	84.70±0.20	91.29±0.16
MIDA-EEG	53.49±0.54	71.67±0.37	41.71±0.83	59.03±0.51	56.94±0.69	73.08±0.46
TCA-EEG	54.72±0.73	72.74±0.69	45.45±1.28	64.60±0.84	59.45±0.74	75.49±0.57
DEEP-Transfer	51.32±0.43	70.93±0.35	46.31±0.39	65.47±0.31	57.65±0.45	73.88±0.52
RA-MDRM	57.25±0.87	74.11±0.59	43.97±1.36	62.36±1.05	60.58±0.52	74.79±0.47
A-Gem	42.73±0.68	62.36±0.45	40.24±0.74	58.80±0.81	53.47±0.43	71.28±0.29
HAL	45.41±3.65	66.05±3.93	46.58±4.66	67.32±3.81	57.93±2.70	74.50±1.82
Meta + ER	56.37±0.83	70.54±0.69	49.21±0.96	71.18±0.64	65.74±0.58	78.66±0.43
Meta + ER + GMED	57.25±0.37	72.98±0.44	51.63±0.98	68.52±0.60	66.29±0.77	79.14±0.46
Meta + MIR	57.68±2.35	71.65±2.11	52.51±3.49	67.74±2.48	68.50±1.83	80.31±1.34
Meta + MIR + GMED	57.02±2.18	73.18±1.70	51.75±1.46	70.29±1.15	69.11±1.27	80.84±0.83
AMBM	59.50±1.33	76.47±0.92	54.93±1.79	72.66±1.33	72.61±0.96	82.78±0.65

Note: The accuracy and ROC-AUC are evaluated immediately after sequential learning of each subject.

stride is 100 for adjacent segments, resulting in 472 segments per trial.

2) Model Settings: The model is implemented with Pytorch and runs on a single TITAN-V GPU. The feature extractor is a three layer CNN network. The first layer performs temporal convolution to generate features with frequency information, and the second layer performs depthwise convolution with spatial filters. Zero padding is applied after each layer to keep original data dimension. One layer of pointwise convolution is appended on top serving as subject-specific prediction head which has reduced computational cost compared to classic convolutional layers. We set filter size of first conv layer to be $(1, C)$, with C the number of underlying channels. The filter size for second conv layer are fixed to be $(2, 32)$ for all three datasets. We keep $k = 6$ prediction heads as default during online training for bi-level adaptation, and the last prediction head is used for backward testing. The hyperparameter η is set to 1e–3, and we performed clipping of –20 and 20 on the gradient of learning rates, which helps stabilize the learning process. A small memory buffer storing 200 data segments is maintained by default, the memory cost of which is negligible compared to the dataset size. With the assumption that training data from the subjects is not available beforehand, we do not use pretrained feature extractor and the model learns sequential arriving subjects on the fly.

E. Analysis on Forward Adaptation

Table II shows the performance of different models for fast adaptation on current subjects, with the accuracy and AUC-ROC evaluated immediately after learning on a subject. We observed the proposed method significantly outperforms comparison models by at least 1.82% on accuracy and 2.36% on AUC-ROC for BCI IV-2a dataset. DEAP dataset is more challenging in general. The proposed model has a significant improvement of 3.18% on accuracy and 1.48% on AUC-ROC. Please note that although the improvement is significant, the overall performance is still low and further improvement is needed. The adaptation task is easier for SEED dataset, with more than half of the models reaching accuracy above 60% during the sequential learning. The proposed model has a 3.50% improvement on accuracy and 1.94% improvement on AUC-ROC compared to best performing counterparts. We conducted detailed examination of performance for the individual decoding tasks in Fig. 4. It reveals the relative improvement of F1-score for the individual decoding tasks in terms of forward adaptation, where F1-score is evaluated immediately after sequential learning of each subject. We observed the AMBM model outperforms other comparison methods by a larger margin on specific tasks, e.g., an improvement of 5.07% on detection of tongue movement and 3.81% on classification of HAPV emotions etc. The result demonstrates the proposed model's ability of fast adaptation during online sequential decoding.

F. Analysis on Forget Mitigation

Table III shows the result of different models toward forgetting mitigation. Here, we evaluate the accuracy on all subjects at the end of sequential learning and also BWT which measures the backward transfer as defined in previous section. The proposed model significantly outperforms these strong comparison models. In particular, our methods outperform other approaches by at least 4.10% on BCI IV-2a, 3.95% on DEAP dataset and 6.19% on SEED dataset in terms of accuracy. The gain on BWT is also significant, with 1.85% on BCI IV-2a dataset, 3.01% on DEAP and 1.86% on SEED dataset. Fig. 5 reveals the relative improvement of F1-score for the individual decoding classes, which is evaluated after all subjects finished sequential learning. The proposed model has a margin of at least 14.28% for tongue movement recognition and 9.45% for the classification of HANV emotions. In general, the model shows higher sensitivity toward high arousal brain activity and the recognition of negative emotions. Fig. 6 depicts the performance of individual subjects and reveals the overall trend of model performance

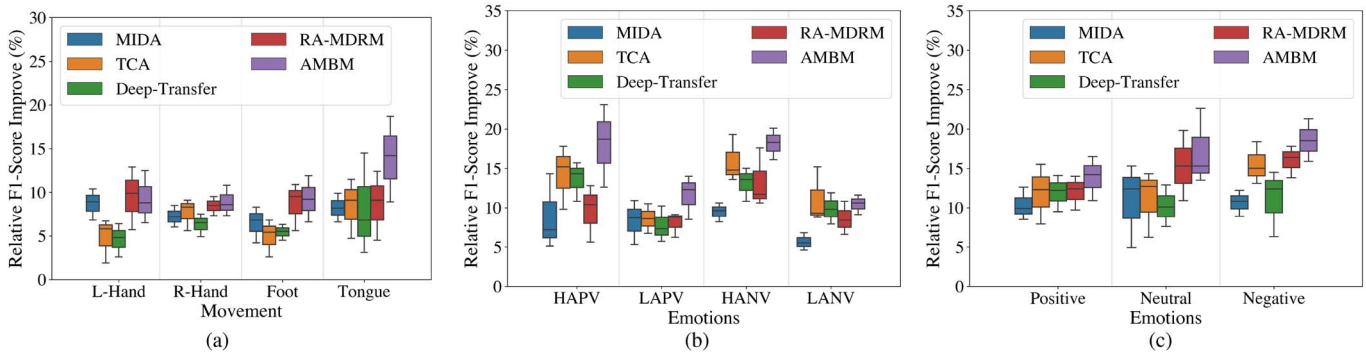

Fig. 4. Adaptation performance improvement relative to base decoder on individual classes in terms of F1-score. (a) BCI IV-2a dataset. (b) DEAP dataset. (c) SEED dataset. The F1-score of individual classes is evaluated immediately after learning of each subject. In (b), HA/LA denotes the high/low arousal, and PV/NV denotes positive/negative valence.

TABLE III
PERFORMANCE ON FORGETTING MITIGATION

Dataset	BCI IV-2a		DEAP		SEED	
Method	Accuracy	BWT	Accuracy	BWT	Accuracy	BWT
Sequential	35.83±0.36	−13.86±0.27	29.31±0.75	−14.90±0.58	36.85±0.43	−17.33±0.31
Joint training	-	-	-	-	-	-
MIDA-EEG	32.47±0.37	−23.61±0.32	30.24±0.59	−12.96±0.44	40.55±0.35	−18.58±0.26
TCA-EEG	31.71±0.39	−25.92±0.41	28.67±0.39	−18.93±0.50	40.87±0.27	−20.96±0.42
DEEP-Transfer	34.69±0.56	−18.84±0.31	31.51±0.68	−16.70±0.39	42.66±0.32	−16.94±0.27
RA-MDRM	30.32±1.20	−30.38±0.73	29.44±1.19	−16.35±0.96	39.78±0.74	−23.49±0.52
A-Gem	34.11±0.46	−9.74±0.53	31.62±0.68	−9.77±0.41	41.73±0.35	−13.26±0.28
HAL	38.57±2.79	−7.81±2.44	33.39±3.61	−14.90±3.22	43.82±3.48	−15.96±2.30
Meta + ER	44.95±0.28	−12.85±0.41	41.70±0.56	−8.41±0.37	54.67±0.49	−12.54±0.23
Meta + ER + GMED	46.32±0.43	−12.34±0.35	42.59±0.73	−10.21±0.66	56.38±0.51	−11.17±0.69
Meta + MIR	49.83±1.48	−9.96±1.29	45.81±2.31	−7.30±2.87	57.22±1.94	−12.78±1.72
Meta + MIR + GMED	50.25±1.37	−7.78±1.60	46.44±2.85	−8.27±2.47	59.52±1.76	−10.80±1.98
AMBM	54.35±1.02	−5.93±1.89	50.49±1.58	−5.26±2.13	65.71±1.49	−8.94±1.70

Note: Evaluation of accuracy and BWT are performed on all subjects at the end of sequential learning.

as sequential decoding progresses. We observed the performance tend to improve for later subjects during the process, as the catastrophic forgetting issue is more severe for earlier subjects and the model retains relatively richer knowledge on later subjects.

G. Performance With Subject Agnostic Setting

We perform experiments toward the more challenging subject-agnostic setting. The result is presented in Table IV. We observed the performance is slightly deteriorating compared to the subject-aware setting. The proposed method has an improvement of 2.63% on BCI IV-2a dataset, 1.75% on DEAP dataset and 5.73% on SEED dataset in terms of accuracy. We observed some comparison methods are on par or slightly better than AMBM in terms of BWT. This is expected as BWT is known to be noisy given it is influenced by both forward adaptation and forget mitigation, and lower accuracy on forward adaptation can result in a better BWT score. We performed detailed analysis of the proposed subject shift detection algorithm in ablation study.

H. Ablation Study

1) Influence From Size of Memory Buffer: We explored the influence of memory buffer size on the proposed method and other memory based continual learning methods including ER, ER + GMED, MIR, MIR + GMED, the result is shown in Table V. The performance on forgetting mitigation is improved with larger memory size. We set memory size to be 200 by default.

2) Subject Shift Detection: Here, we explore the performance of the proposed subject shift detection algorithm for the three datasets. The number of shifts is number of subjects minus 1, and for each dataset we run the experiment repetitively for 10 times. We provide visualization on the probabilistic distribution of run length in Fig. 7. The probability of run length dropping to zero is consistent with the majority cases of subject shifts. The algorithm successfully detected 63 out of the 80 occurrences for BCI IV-2a, 225 out of the 310 occurrences for DEAP dataset, 122 out of the 140 occurrences for SEED dataset, rendering detection accuracy of 78.8% for BCI IV-2a, 72.6% for DEAP and 87.1% for SEED dataset.

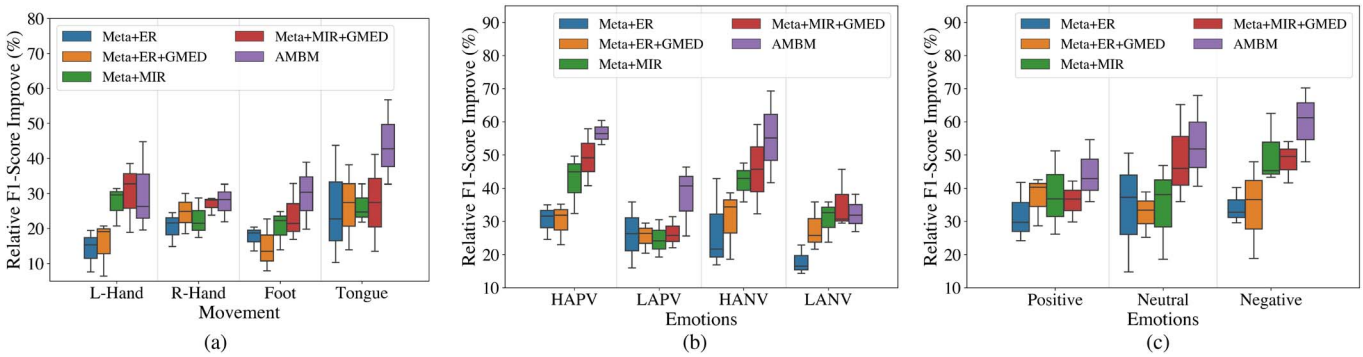

Fig. 5. Forgetting mitigation improvement relative to base decoder on individual classes in terms of F1-score. Here the F1-score of individual classes is evaluated at the end after all subjects finished sequential learning. (a) BCI IV-2a dataset. (b) DEAP dataset. (c) SEED dataset.

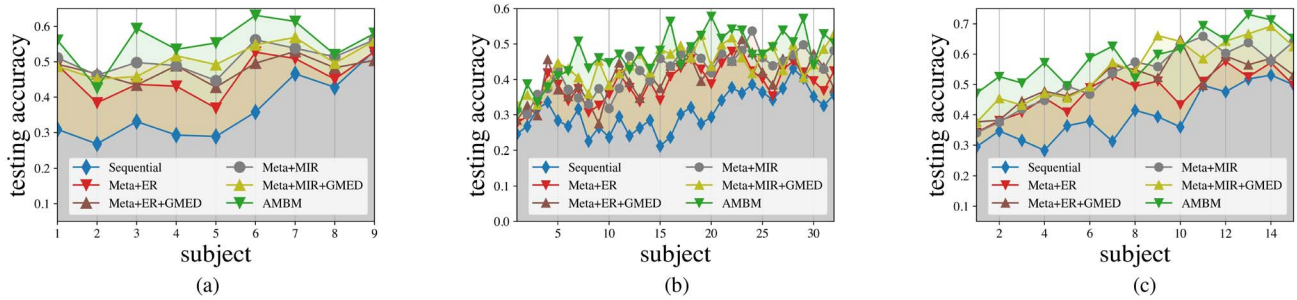

Fig. 6. Performance of individual subjects at the end of sequential learning. The AMBM method enables the model to achieve decent performance on earlier subjects. (a) BCI IV-2a dataset. (b) DEAP dataset. (c) SEED dataset.

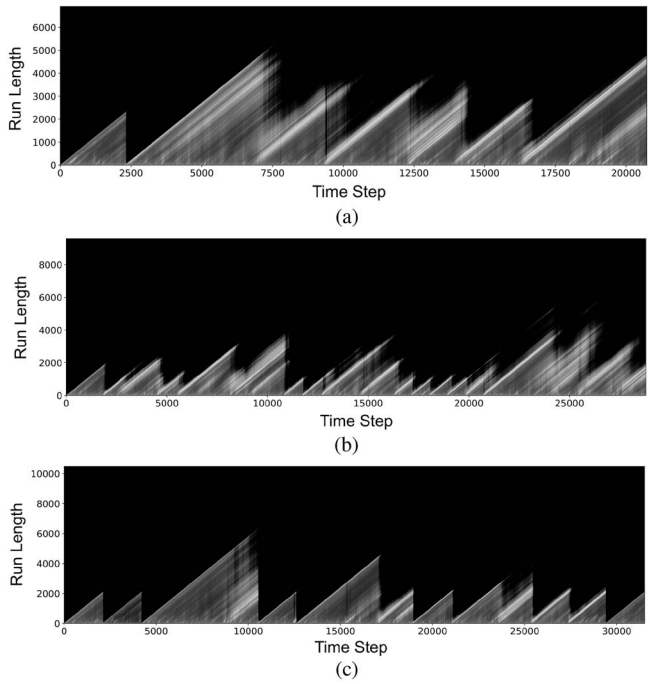

Fig. 7. Visualization on posterior distribution $P(l_{S_t}|\boldsymbol{d}_{1:t})$ during subject shift detection, with x axis being the time step and y axis being the run length. (a) BCI IV-2a dataset. (b) DEAP dataset. (c) SEED dataset.

We performed further analysis on our distance metric based approach compared to direct usage of embedding from feature extractor, based on the maximal information coefficient (MIC) [54] and total information coefficient(TIC) [55] for evaluation of correlation before and after the subject shift happens. These two metrics are indicators of nonlinear dependencies between variables. A smaller value means the variables are more independent of each other, which is beneficial for the variable to be used for subject shift detection as it filters out the overlapping mutual information and makes it easier for the task. The result is shown in Table VI. We observed the distance metric based $\boldsymbol{d}_t$ is significantly more independent of each other across different time steps compared to direct usage of feature embedding $\boldsymbol{e}_t$. This illustrates the privilege of extracting distance metric information from raw embeddings, which filters out the correlated semantic information of the signal and reduces the overlapping mutual information.

3) Module Ablation Settings: We performed a comprehensive comparison on the different ablation settings of the proposed modules, with the result summarized in Table VII. We observed the inclusion of mutual information maximization across subjects in meta loop (MiM) contributes not only to the improved performance of forward adaptation, but also achieves significant performance gains for forgetting mitigation with improved generalization across different subjects. The holistic feature extraction with mutual information maximization in base loop (MiB) show benefits on the forward adaptation process while the improvement on knowledge retaining is relatively modest. On the other hand, the dynamic hyperparameter adaptation (HA) achieved notable gains for forgetting mitigation

TABLE IV
PERFORMANCE WITH SUBJECT AGNOSTIC SETTING

Dataset	BCI IV-2a		DEAP		SEED	
Method	Accuracy	BWT	Accuracy	BWT	Accuracy	BWT
Meta + ER	44.95±0.28	−12.85±0.41	41.70±0.56	−8.41±0.37	54.67±0.49	−12.54±0.23
Meta + ER + GMED	46.32±0.43	−12.34±0.35	42.59±0.73	−10.21±0.66	56.38±0.51	−11.17±0.69
Meta + MIR	49.83±1.48	−9.96±1.29	45.81±2.31	−7.30±2.87	57.22±1.94	−12.78±1.72
Meta + MIR + GMED	50.25±1.37	−7.78±1.60	46.44±2.85	−8.27±2.47	59.52±1.76	−10.80±1.98
AMBM	52.88±1.73	−7.72±2.27	48.19±2.10	−7.84±2.36	65.25±1.24	−9.53±1.48

Note: The AMBM model functions on proposed subject shift detection algorithm. We observed the performance is slightly deteriorating compared to the subject aware setting.

TABLE V
INFLUENCE OF MEMORY SIZE ON MODEL PERFORMANCE

Dataset	BCI IV-2a			DEAP			SEED		
Memory Size	100	200	500	100	200	500	100	200	500
Meta + ER	38.52±0.58	44.95±0.73	56.48±0.37	35.31±0.64	41.70±0.85	50.64±0.59	48.47±0.38	54.63±0.44	63.51±0.25
Meta + ER + GMED	41.63±0.45	46.39±0.60	57.96±0.32	37.24±0.74	42.58±0.66	52.42±0.39	50.89±0.41	56.35±0.58	65.17±0.27
Meta + MIR	45.73±1.34	49.84±1.68	59.39±0.85	39.51±2.73	45.80±3.09	54.26±2.16	51.34±3.33	57.28±2.75	66.64±2.04
Meta + MIR + GMED	45.95±1.85	50.22±2.47	59.14±1.53	40.79±3.70	46.41±2.44	55.97±1.91	51.95±3.20	59.56±2.50	67.36±2.73
AMBM	48.27±0.86	54.35±1.24	61.76±2.18	44.10±2.32	50.39±1.58	58.64±1.85	55.61±1.49	65.71±1.49	70.84±0.83

Note: The performance of forgetting mitigation is improved with larger memory size.

TABLE VI
NON-LINEAR CORRELATION TEST OF THE PROPOSED DISTANCE METRIC APPROACH COMPARING TO DIRECT USAGE OF FEATURE EMBEDDING FOR SUBJECT SHIFT DETECTION

Dataset	BCI IV-2a		DEAP		SEED	
Non-linear Correlation	MIC	TIC	MIC	TIC	MIC	TIC
Feature embedding $\boldsymbol{e}_t$	0.3348±0.0043	0.2133±0.0037	0.2919±0.0025	0.1962±0.0031	0.3493±0.0044	0.2274±0.0020
Distance metric $\boldsymbol{d}_t$	0.1931±0.0017	0.1057±0.0026	0.2045±0.0041	0.1183±0.0012	0.1975±0.0038	0.0981±0.0024

TABLE VII
DETAILED COMPARISON ON THE DIFFERENT MODULE ABLATION SETTINGS

Dataset	BCI IV-2a		DEAP		SEED	
Modules	Adaptation	Retaining	Adaptation	Retaining	Adaptation	Retaining
Base Model	48.12±0.24	35.83±0.36	42.56±0.49	29.31±0.75	52.28±0.39	36.85±0.43
Meta + MiB	55.30±0.87	37.51±0.62	47.72±1.15	33.26±1.40	64.23±0.54	42.51±0.96
Meta + MiM	58.73±1.44	46.29±0.95	52.38±0.74	43.85±1.03	69.40±1.27	56.94±0.70
Meta + HA	53.41±1.28	49.62±0.76	48.33±0.67	46.59±1.12	64.61±0.89	60.32±0.58
Meta + MiB + MiM	60.26±2.05	45.17±1.64	54.61±1.32	45.90±0.86	73.25±1.42	58.07±1.13
Meta + MiB + HA	56.58±0.61	51.05±1.47	49.24±0.96	46.44±1.54	66.82±0.93	61.60±2.08
Meta + MiM + HA	58.97±1.19	52.61±0.83	53.15±1.28	50.21±0.62	70.33±1.71	64.82±0.95
Meta + MiB + MiM + HA	59.50±1.33	54.35±1.02	54.93±1.79	50.49±1.58	72.61±0.96	65.71±1.49

Note: "MiB" denotes the mutual information maximization in base loop, "MiM" denotes the mutual information maximization in meta loop and "HA" refers to the dynamic hyperparameter adaptation. We evaluated both the performance for subject adaptation (evaluated immediately after learning of each subject) and knowledge retaining (evaluation performed on all subjects at the end of sequential learning).

while having minor influence on subject adaptation. In Fig. 8, we visualized the ratio of improvement relative to the margin of AMBM model (compared to base decoder) for the different module ablation settings.

4) Settings Toward Mutual Information Maximization: Here, we study the different settings in mutual information maximization for online adaptation of subjects. The model computes mutual information on features of both current data and memory data in meta loop. Two different approaches exist for this, 1) perform feature concatenation and then perform MI computation, which we denote as $I_{\{\mathcal{F}_t,\mathcal{F}_m\}}$; and 2) perform addition of mutual information at the end, which we denote

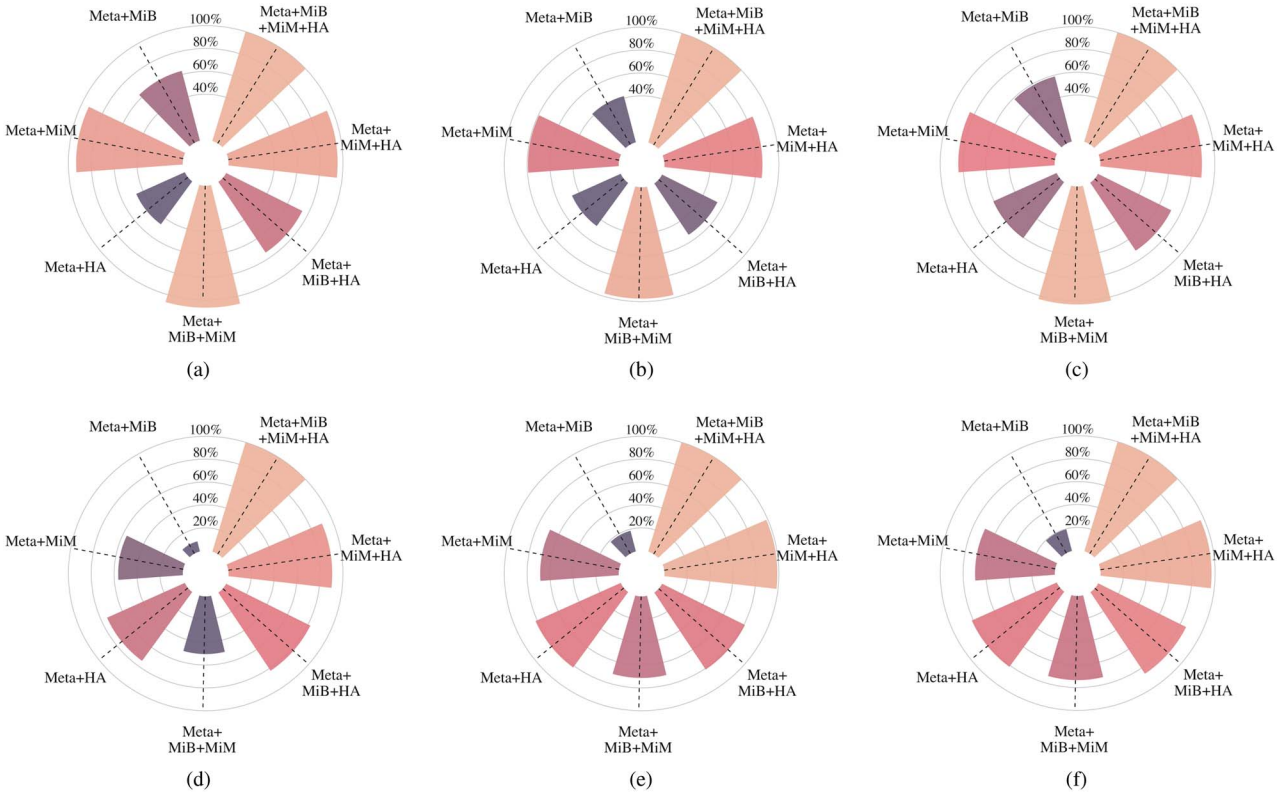

Fig. 8. Visualization on the ratio of improvement relative to the margin of AMBM model (compared to base decoder) for the different ablation module settings. The rows from top to bottom corresponds to subject adaptation and knowledge retaining performance respectively. The columns corresponds to the three different datasets including BCI IV-2a, DEAP and SEED. (a) Adaptation on BCI IV-2a. (b) Adaptation on DEAP. (c) Adaptation on SEED. (d) Retaining on BCI IV-2a. (e) Retaining on DEAP. (f) Retaining on SEED

TABLE VIII
MODEL PERFORMANCE WITH DIFFERENT MI SETTINGS

Dataset	BCI IV-2a			DEAP			SEED		
MI Setting	$I_{\{\mathcal{F}_t\}\{\mathcal{F}_m\}}$	$I_{\mathcal{F}_t}$	$I_{\{\mathcal{F}_t,\mathcal{F}_m\}}$	$I_{\{\mathcal{F}_t\}\{\mathcal{F}_m\}}$	$I_{\mathcal{F}_t}$	$I_{\{\mathcal{F}_t,\mathcal{F}_m\}}$	$I_{\{\mathcal{F}_t\}\{\mathcal{F}_m\}}$	$I_{\mathcal{F}_t}$	$I_{\{\mathcal{F}_t,\mathcal{F}_m\}}$
MI + ER	51.47±0.63	49.52±0.45	52.86±0.72	47.81±0.53	45.93±0.86	50.40±0.39	62.67±0.47	62.32±0.33	63.79±0.60
MI + MIR	52.10±2.48	50.97±1.34	55.25±1.51	49.58±1.26	46.21±2.07	51.34±1.66	63.59±1.34	62.83±2.58	64.16±2.02
AMBM	54.35±1.24	53.74±0.89	56.48±1.07	50.39±1.58	48.63±0.61	52.12±0.95	65.71±1.49	64.24±1.32	66.75±0.91

Note: Concatenating extracted features of current data $\mathcal{F}_t$ and memory data $\mathcal{F}_m$ before the MI computation renders best performance among the three different settings.

TABLE IX
MODEL PERFORMANCE WITH DIFFERENT SUBJECT ORDERING

Dataset	BCI IV-2a			DEAP			SEED		
Ordering	Sequential	Ascending	Descending	Sequential	Ascending	Descending	Sequential	Ascending	Descending
A-Gem	34.11±0.46	36.73±0.71	33.67±0.36	31.62±0.68	32.96±1.20	29.21±0.44	41.73±0.35	44.24±0.47	39.58±0.72
HAL	38.57±2.79	39.68±3.65	38.15±2.36	33.39±3.61	34.13±1.48	30.86±3.32	43.82±3.48	46.13±2.06	42.96±3.14
Meta + ER	44.95±0.28	46.11±0.43	43.60±0.62	41.70±0.56	43.73±0.78	39.57±0.31	54.67±0.49	56.82±0.63	51.79±0.24
Meta + ER + GMED	46.32±0.43	48.54±0.76	45.17±0.49	42.59±0.73	44.28±0.52	40.76±0.60	56.38±0.51	57.97±0.87	54.41±0.63
Meta + MIR	49.83±1.48	50.45±2.53	48.81±1.76	45.81±2.31	47.14±2.81	42.98±1.63	57.22±0.94	59.04±2.26	54.99±1.85
Meta + MIR + GMED	50.25±1.37	51.81±1.56	49.36±2.10	46.44±2.85	48.66±1.68	45.21±1.39	59.52±1.76	61.36±2.62	58.13±2.41
AMBM	54.35±1.24	55.75±1.43	53.23±1.85	50.39±1.58	51.54±1.71	48.61±1.93	65.71±1.49	67.47±0.86	63.92±1.15

Note: We conducted ablation study on three ordering settings: 1) ascend ordering based on id of subjects; 2) ascend ordering based on decoding accuracy; and 3) descend

as $I_{\{\mathcal{F}_t\}\{\mathcal{F}_m\}}$. And we also compare to only performing the MI operation on features of current data $\mathcal{F}_t$. Table VIII provides details on the comparison. We observed among the three settings, the performance is worst when MI is only applied to $\mathcal{F}_t$. The concatenation of features $\mathcal{F}_t$ and $\mathcal{F}_m$ renders better performance compared to adding MI outputs in the end.

TABLE X
INFLUENCE OF CONTEXT SIZE ON MODEL PERFORMANCE FOR DOMAIN AGNOSTIC SETTING, ON ALL THREE DATASETS

Context Size	2	4	6	8	10
Dataset	ACC				
BCI IV-2a	50.17±0.84	50.46±1.32	51.38±1.73	51.10±0.76	50.81±1.07
DEAP	46.15±1.53	46.44±1.89	46.96±1.29	47.19±2.10	46.54±1.48
SEED	63.76±0.87	64.24±1.62	64.65±1.24	64.38±1.45	63.91±1.06

5) Effect of Subject Ordering on Model Performance: The level of difficulty in decoding differs across the subjects. For example, subjects A02 and A04 are more challenging to decode than A01 and A03 in BCI IV-2a dataset. Here, we explore the different ordering of subjects toward model performance. We study three different ordering scenarios, the first is sequential order, the second is ascending order based on eval accuracy, the third one is descending order based on eval accuracy. We summarized the result in Table IX. The ascending order is achieving the best performance among the three scenarios while the descending order is most difficult to decode. This reveals that forgetting is more severe when sequential learning begins with easier subjects and ends with more challenging ones.

6) Window Size of Distance Metric: The window size for distance metric calculation in Section III-D determines the range of past time steps that is incorporated into consideration. We performed ablation study on the effect of this window size toward performance. The result is provided in Table X. We found there is a tradeoff between window size and model performance. For window size increasing from 2 to 10, the performance improves at the beginning and then peaks and deteriorates with larger size.

V. DISCUSSION AND CONCLUSION

In this work, we proposed a principled approach for online sequential decoding of brain signals with the occurrence of subject shifts, and simultaneously tackles the challenges of fast adaptation and forgetting mitigation. The proposed method forms a meta optimizer with bi-level mutual information maximization for fast adaptation on incoming subjects, which enables effective alignment across different subjects and significantly improves the generalization ability for fast adaptation purposes. At the same time, the adaptive hyperparameter update at meta loop effectively releases the potential of different portions of the neural network in retaining previously learned knowledge and alleviates the catastrophic forgetting issue. We formed challenging benchmarks for online sequential EEG decoding on top of three different public EEG datasets. The extensive empirical evaluation reveals the privilege of proposed method, outperforming strong baselines in numerous different ablation scenarios.

The limitations of the work include the following: 1) the model maintains a small memory (for both subject alignment and hyperparameter adaptation purposes), this increases the memory cost to be slightly higher than base EEG decoder (meanwhile comparable to continual learning baselines such as ER and MIR); 2) Though the model achieved relatively large improvement compared to baselines for the online sequential decoding scenario, there is still a large performance gap from the upper bound, further performance improvement is needed.

The model has wide applicability in real-world scenarios such as robotic wheelchair control and rehabilitation training, with new subjects arriving sequentially. Specifically, 1) fast adaptation enables the model to achieve good performance on newly arriving subjects; and 2) forgetting mitigation frees the model from repeated training every time a new subject arrives, instead it just needs to learn on the new subject and previous knowledge is effectively retained by the model.

There are many directions for exploration of future work, including: 1) exploration on sequence of subjects with heterogeneous classes; 2) thorough analysis on the correlation of decoding accuracy with similarities of adjacent subjects; and 3) methods for online sequential decoding without usage of memory buffer.

REFERENCES

[1] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, pp. 1–34, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/6/1423

[2] L. Bi, X.-A. Fan, and Y. Liu, "EEG-based brain-controlled mobile robots: A survey," *IEEE Trans. Hum. Mach. Syst.*, vol. 43, no. 2, pp. 161–176, Mar. 2013.

[3] V. K. K. Shivappa, B. Luu, M. Solis, and K. George, "Home automation system using brain computer interface paradigm based on auditory selection attention," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 1–6.

[4] C. Gómez, P. Arbeláez, M. Navarrete, C. Alvarado-Rojas, M. Le Van Quyen, and M. Valderrama, "Automatic seizure detection based on imaged-EEG signals through fully convolutional networks," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020.

[5] R. Salazar-Varas and R. A. Vazquez, "Facing high EEG signals variability during classification using fractal dimension and different cutoff frequencies," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–12, May 2019.

[6] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 2390–2397.

[7] J. Fdez, N. Guttenberg, O. Witkowski, and A. Pasquali, "Cross-subject EEG-based emotion recognition through neural networks with stratified normalization," *Front. Neurosci.*, vol. 15, pp. 1–14, Feb. 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.626277

[8] P. Pandey and K. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1730–1738, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1319157819309991

[9] T. Duan et al., "Meta learn on constrained transfer learning for low resource cross subject EEG classification," *IEEE Access*, vol. 8, pp. 224791–224802, 2020.

[10] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination," *Front. Neurorobotics*, vol. 11, pp. 1–16, Apr. 2017. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2017.00019

[11] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *J. Neural Eng.*, vol. 16, no. 2, Jan. 2019, Art. no. 026007. doi: 10.1088/1741-2552/aaf3f6

[12] S. Siuly, Y. Li, and Y. Zhang, *EEG Signal Analysis and Classification: Techniques and Applications*, 1st ed. Cham, Switzerland: Springer-Verlag, 2018.

[13] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, "Real-time EEG-based happiness detection system," *Sci. World J.*, vol. 2013, Aug. 2013, Art. no. 618649.

[14] D. Wu, B. J. Lance, V. J. Lawhern, S. Gordon, T. Jung, and C. Lin, "EEG-based user reaction time estimation using Riemannian geometry

features," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 2157–2168, Nov. 2017.
[15] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: A review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1753–1762, Oct. 2017.
[16] F. Farkhondeh Tale Navi, S. Heysieattalab, D. S. Ramanathan, M. R. Raoufy, and M. A. Nazari, "Closed-loop modulation of the self-regulating brain: A review on approaches, emerging paradigms, and experimental designs," *Neuroscience*, vol. 483, pp. 104–126, Aug. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306452221006217
[17] M. Shanechi, "Brain–machine interfaces from motor to mood," *Nat. Neurosci.*, vol. 22, pp. 1554–1564, Oct. 2019.
[18] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Jul. 2018, Art. no. 056013.
[19] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730
[20] D. Zhang et al., "Cascade and parallel convolutional recurrent neural networks on EEG-based intention recognition for brain computer interface," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–7.
[21] M. D. Lange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.
[22] M. Riemer et al., "Learning to learn without forgetting by maximizing transfer and minimizing interference," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–31.
[23] A. Chaudhry et al., "On tiny episodic memories in continual learning," 2019, *arXiv:1902.10486*.
[24] R. Aljundi et al., *Online Continual Learning with Maximally Interfered Retrieval*. Red Hook, NY, USA: Curran Associates Inc., 2019.
[25] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS'17)*, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6470–6479. [Online]. Available: http://dl.acm.org/citation.cfm?id=3295222.3295393
[26] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–20.
[27] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci.*, pp. 1–15, Apr. 2017.
[28] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.
[29] S. C. Y. Hung, C.-H. Tu, C.-E. Wu, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, *Compacting, Picking and Growing for Unforgetting Continual Learning*. Red Hook, NY, USA: Curran Associates Inc., 2019.
[30] A. A. Rusu et al., "Progressive neural networks," 2016, *arXiv: 1606.04671*. [Online]. Available: http://arxiv.org/abs/1606.04671
[31] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–11.
[32] M. Zhou et al., "Image de-raining via continual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 4905–4914.
[33] H. Chawla, A. Varma, E. Arani, and B. Zonooz, "Continual learning of unsupervised monocular depth from videos," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8419–8429.
[34] Z. Ke, H. Xu, and B. Liu, "Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2021, pp. 1–10. [Online]. Available: https://api.semanticscholar.org/CorpusID:235097247
[35] Y. Zhang, X. Wang, and D. Yang, "Continual sequence generation with adaptive compositional modules," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Volume 1 Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3653–3667. [Online]. Available: https://aclanthology.org/2022.acl-long.255
[36] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.
[37] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *Adv. Neural Inf. Process. Syst.*, pp. 1–22, Jan. 2017.
[38] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–10.
[39] A. A. Rusu et al., "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=BJgklhAcK7
[40] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, "Online meta-learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1–11.
[41] L. Li et al., "Progressive domain expansion network for single domain generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)* Virtual, June 19–25, 2021, Piscataway, NJ, USA: IEEE Press, 2021, pp. 224–233.
[42] Y. Guo, B. Liu, and D. Zhao, "Online continual learning through mutual information maximization," in *Proc. 39th Int. Conf. Mach. Learn.*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, Sep. 2022, pp. 1–8. [Online]. Available: https://proceedings.mlr.press/v162/guo22g.html
[43] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[44] M. Tangermann et al., "Review of the BCI competition IV," *Front. Neurosci.*, vol. 6, Oct. 2012, Art. no. 55.
[45] S. Koelstra et al., "DEAP: A database for emotion analysis;using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.
[46] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[47] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER),* Piscataway, NJ, USA: IEEE Press, 2013, pp. 81–84.
[48] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI'16)*, Washington DC, USA: AAAI Press, 2016, pp. 2732–2738.
[49] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cogn. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.
[50] C. Tan, F. Sun, and W. Zhang, "Deep transfer learning for EEG-based brain computer interface," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 916–920.
[51] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.
[52] A. Chaudhry, A. Gordo, P. K. Dokania, P. H. S. Torr, and D. Lopez-Paz, "Using hindsight to anchor past knowledge in continual learning," 2020, *arXiv:2002.08165*.
[53] X. Jin, A. Sadhu, J. Du, and X. Ren, "Gradient based memory editing for task-free continual learning," 2020, *arXiv:2006.15294*.
[54] D. N. Reshef et al., "Detecting novel associations in large data sets," *Sci.*, vol. 334, no. 6062, pp. 1518–1524, 2011.
[55] Y. A. Reshef, D. N. Reshef, H. K. Finucane, P. C. Sabeti, and M. Mitzenmacher, "Measuring dependence powerfully and equitably," *J. Mach. Learn. Res.*, vol. 17, no. 211, pp. 1–63, 2016.