Replay with Stochastic Neural Transformation for Online Continual EEG Classification

Tiehang Duan[1,3], Zhenyi Wang[2], Gianfranco Doretto[3], Fang Li[4], Cui Tao[4], Donald A. Adjeroh[3]
[1] Meta AI, Bellevue, WA 98005, United States
[2] University of Maryland, College Park, MD 20742, United States
[3] West Virginia University, Morgantown, WV 26506, United States
[4] University of Texas Health Science Center at Houston, Houston, TX 77030, United States
tiehang.duan@gmail.com, wangzhenyineu@gmail.com, gianfranco.doretto@mail.wvu.edu
fang.li@uth.tmc.edu, cui.tao@uth.tmc.edu, donald.adjeroh@mail.wvu.edu

***Abstract*—Brain computer interface (BCI) systems used for clinical assistance purposes such as wheelchair control require decoding of streaming brain signals i.e. electroencephalography (EEG) signals over a long period of time with subject shift in the middle. Numerous challenges arise during this online continual brain signal decoding process: 1) the EEG decoder needs to deal with streaming EEG signals from sequentially arriving subjects, with no data available beforehand for large-scale pretraining; 2) the EEG decoder should avoid catastrophic forgetting on previous subjects after learning on a new subject; 3) the EEG decoder should perform well on noisy signals with high variance across subjects. We proposed a principled replay-based approach for this general decoding scenario, forming a bi-level optimization framework with stochastic neural transformation for dynamic memory evolution, making them representative in feature space and encouraging the model to generalize well. The evolved signal segments are stored and replayed during later decoding stages to achieve optimal model performance on all previous subjects. The stochastic neural transformation performed in inner sup of bi-level optimization significantly enhances the diversity of stored signal segments and improves model robustness during online continual decoding. We perform detailed theoretical analysis on model's generalization ability in addition to the empirical evaluations. We construct multiple new benchmarks to mimic real-world online sequential EEG decoding scenarios with underlying subject shifts. The extensive evaluation of the proposed approach shows it outperforms related strong baselines by a large margin.**

I. Introduction

Development of brain computer interface (BCI) systems for clinical assistance purposes such as robotic wheelchair control or digital interface interaction requires proper decoding of electroencephalography (EEG) signals for a long period of time [1], [2]. Given the non-invasive nature of EEG recording, the signal usually endures significant noise and its patterns differ significantly across different subjects. This poses challenge to the BCI system when it needs to sequentially decode streaming EEG signals from different subjects.

Here, We formulate the problem as that of online continual EEG decoding. The problem setting is characterized by the following challenges: 1) the model decodes streaming EEG signals in an online manner and no data is available for pre-training; 2) the streaming signals are from sequentially arriving subjects with subject shifts; 3) the EEG decoder should not forget knowledge of previous subjects after learning new ones, aka. catastrophic forgetting, and achieve optimal performance on all subjects after sequential decoding ends[1].

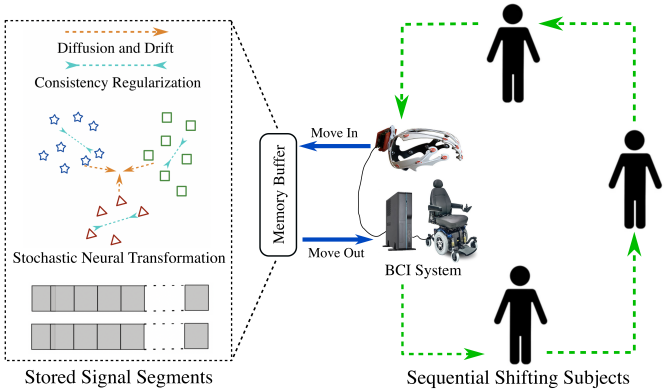


Fig. 1: An overview of the replay-based ReSNT model for online continual decoding of streaming EEG signals over a long period of time with subjects shift. We propose to perform stochastic neural transformation on the stored segments following the bi-level optimization target (eq. 6-7) to help the model generalize to all previous subjects.

Deep learning approaches have demonstrated effectiveness for EEG decoding in general and achieved state-of-the-art performance [3][4]. Prior efforts have focused mostly on developing new model architectures [5][3][6] or exploring domain adaptation/transfer learning techniques [7][8] for classic EEG decoding settings e.g. cross-subject EEG decoding, etc. To the best of our knowledge, little exploration has been made for the challenging online continual EEG decoding problem. Recently, continual learning techniques emerged to be promising for mitigation of the problem of catastrophic forgetting [9], [10], with approaches including replay based methods [11], regularization based methods [12] and dynamic expandable network architectures [13]. In this work, we focus on replay-based approaches for continual decoding

[1]Classic cross subject EEG decoding models are not readily applicable to this problem setting, as the signal is streaming in an online manner and the data is not jointly available beforehand for large scale pre-training.

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

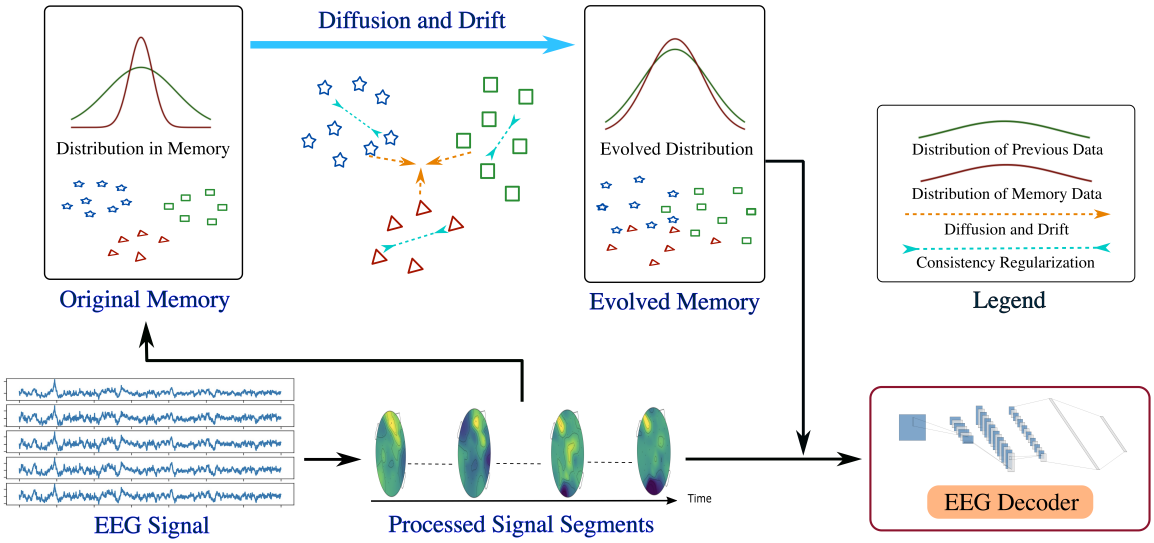


Fig. 2: Illustration on the evolution of signal segments following the stochastic neural transformation. The divergence and drift operations produce robust and diversified features for generalization, and the consistency regularization preserves similarity between transformed signal and original signal.

across sequentially arriving different subjects. The proposed approach is significantly different from existing methods, with its emphasis on generalization improvement and overfitting mitigation during online continual EEG decoding.

Specifically, the proposed replay-based approach stores a small number of representative signal segments from previous subjects for replay purpose. Given that the amount of stored segments are very limited, overfitting could be a major problem and cause the model not generalize well on previous data. As explored in previous work, the diversity of training data is a major factor for model's generalization ability [14]. We form a bi-level optimization framework with stochastic neural transformation on signal segments to keep them diverse and representative for replay purposes. The transformation process consists of two aspects: 1) diffusion and drift operations that evolve the memory data to improve generalization ability. 2) consistency regularization that enforces similarity with original data and produces smooth model response, which allows the model to achieve optimal performance on testing data. The neural transformation process can be depicted as stochastic neural ordinary differential equations, which enable the data to go through an infinite number of stochastic transformations and improve data diversity in the process.

The contribution of this work is summarized as follows:

- We propose a replay-based approach with dynamic memory evolution for online continual decoding of streaming EEG signals over a long period of time with subject shifts.
- The proposed approach performs bi-level optimization with memory evolution depicted as a stochastic process. The evolution process is regularized with consistency constraints to preserve similarity with original signal.
- We developed online sequential EEG decoding benchmarks, and performed extensive evaluation on the model performance. The results show the proposed approach outperformed strong baselines by a significant margin.

II. Method

We first offer a high-level description of the problem setting and form the overall target function, then provide details of the proposed online continual decoding approach named **Re**play with **S**tochastic **N**eural **T**ransformation (ReSNT) under the bi-level optimization framework.

A. Preliminaries

For decoding over a long period of streaming EEG signals with subject shift in the middle, decoding performance can quickly deteriorate on previous subjects after learning on later subjects. Here we seek to achieve optimal decoding performance on all previously learnt subjects and mitigate forgetting during online sequential EEG decoding. Note we assume there is no jointly available data from the subjects for pretraining beforehand, which suits to real world scenarios of BCI applications.

We adopt the replay based approach, which keeps a small memory buffer $\mathcal{M}$ to store representative data samples of previous subjects and replay in later decoding stages. With streaming EEG signal from different subjects $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_J$ sequentially input into the EEG decoder, the memory buffer adopts the reservoir sampling mechanism for sample selection, which has an equal probability of selecting incoming signal segments. It stores the first M samples until it is full, after which for data x_k arriving, it generates a random number i in the range of [1, k]. If $i < M$, then the i th data in memory will be replaced by x_k.

B. Method Overview

The training target for the replay based approach can be formulated as

$$\min_{\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[\mathcal{L}(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) + \mathop{\mathbb{E}}_{(\boldsymbol{x}_m, y_m) \sim \mathcal{M}} \mathcal{L}(\boldsymbol{x}_m, y_m, \boldsymbol{\theta}) \right], \quad (1)$$

where $\boldsymbol{x}_i, y_i$ is the current streaming data and $\boldsymbol{x}_m, y_m$ is the stored data in memory. When we only keep a small amount of data in memory, the conventional replay will be less effective when sequentially learning over a longer period of time, as the model would easily overfit onto the limited memory data and may not successfully generalize to previous subjects. Here we utilize the stochastic neural transformation with consistency constraints to achieve improved diversity in stored signal segments for effective replay.

C. Stochastic Neural Transformation during Replay

The proposed stochastic neural transformation creates diversified data with the diffusion and drift operations that depict the evolution of a stochastic dynamic system. With $\sigma_{\boldsymbol{\phi}} : [0, G] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times w}$ being the diffusion term and $\mu_{\boldsymbol{\phi}} : [0, G] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ being the drift term, both generated by a network parameterized with $\boldsymbol{\phi}$ and G is the transformation intensity, the stochastic neural transformation process can be formulated as

$$dX_t = \mu_{\boldsymbol{\phi}}(X_t)dt + \sigma_{\boldsymbol{\phi}}(X_t) \circ dW_t, \tag{2}$$

where X_t is the distribution of transformed data at t, with $t \in [0, G]$. and $X_t \in \mathbb{R}^d$. $W_t \rightarrow \mathcal{W}^w$ is a w-dimensional variable following Brownian motion, with the property that $W_{t+r} - W_r$ follows a Gaussian distribution with variance t. The initial state of the neural transformation is the original stored segments. The transformed data can be expressed as

$$X_G = X_0 + \int_0^G \mu_{\boldsymbol{\phi}}(X_t)dt + \int_0^G \sigma_{\boldsymbol{\phi}}(X_t) \circ dW_t \tag{3}$$

The $\circ$ operation between $\sigma_{\boldsymbol{\phi}}(X_t)$ and dW_t denotes the Stratonovich integral.

In order for the transformed data to be consistent with the original data, we utilize the Jenson-Shannon divergence regularization to produce smooth model response on the transformed data, which functions between the posterior distribution of transformed data and original data.

$$p_{mean} = (p_{\boldsymbol{x}_0} + p_{\boldsymbol{x}_G})/2 \tag{4}$$

$$\mathbb{JS}(\boldsymbol{x}_G, \boldsymbol{x}_0) = (\mathbb{KL}(p_{\boldsymbol{x}_0} || p_{mean}) + \mathbb{KL}(p_{\boldsymbol{x}_G} || p_{mean}))/2 \tag{5}$$

where $\mathbb{KL}$ is the KL divergence, $p_{\boldsymbol{x}_0}$ and $p_{\boldsymbol{x}_G}$ are the network output probability of classifying the original and transformed data into different classes.

Parameters of the stochastic neural transformer $\boldsymbol{\phi}$ is learned in an end-to-end manner together with the model parameters $\boldsymbol{\theta}$ with bi-level optimization. Here, the target loss function can be expressed as

$$\min_{\boldsymbol{\theta}} [\mathcal{L}(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) + \mathcal{L}(\boldsymbol{x}_m(G), \boldsymbol{y}_m, \boldsymbol{\theta}, \boldsymbol{\phi}_*)] \tag{6}$$

$$s.t. \quad \boldsymbol{\phi}_* = \arg\max_{\boldsymbol{\phi}} [\mathcal{L}(\boldsymbol{x}_m(G), \boldsymbol{y}_m, \boldsymbol{\theta}, \boldsymbol{\phi}) - \mathcal{L}(\boldsymbol{x}_m, \boldsymbol{y}_m, \boldsymbol{\theta}, \boldsymbol{\phi}) - \lambda \mathbb{JS}(\boldsymbol{x}_m(G), \boldsymbol{x}_m)] \tag{7}$$

where $\boldsymbol{x}_m(G)$ and $\boldsymbol{x}_m$ are the transformed and original stored segments respectively. The minimization target of eq. 6 performs optimization on model parameter $\boldsymbol{\theta}$, and the stochastic neural transformation parameter $\boldsymbol{\phi}$ is optimized with eq. 7, which diverges the transformed data not to be memorized by model and at the same time being consistent and semantically similar to original data.

To perform proper update on $\sigma_{\boldsymbol{\phi}}(X_t)$ and $\mu_{\boldsymbol{\phi}}(X_t)$ based on the target function specified in eq. 7, we perform the following:

With $A_t = \frac{d\mathcal{L}(X_t)}{dX_t}$, the relationship between A_t and $\sigma_{\boldsymbol{\phi}}(X_t)$, $\mu_{\boldsymbol{\phi}}(X_t)$ can be expressed as

$$dA_t^i = -A_t^j \frac{\partial \mu_{\boldsymbol{\phi}}^j(X_t)}{\partial X^i} dt - A_t^j \frac{\partial \sigma_{\boldsymbol{\phi}}^{j,k}(X_t)}{\partial X^i} \circ dW_t^k \tag{8}$$

where i, j, k are the index of matrix A_t and $\sigma_{\boldsymbol{\phi}}(X_t)$. The gradient on $\boldsymbol{\phi}$ can be obtained accordingly based on eq. 8. Different from conventional data augmentation and evolution approaches which rely on empirical evaluation for the effectiveness to tackle overfitting, the proposed approach offers a principled formulation that enables theoretical analysis on its generalization improvement. Details omitted for brevity.

D. Algorithm

Our overall approach is summarized in Algorithm 1. The learning algorithm alternates between update of EEG decoder paramters $\boldsymbol{\theta}$ and neural transformation parameters $\boldsymbol{\phi}$. We use reversible Heun method [15] to perform update on $\boldsymbol{\phi}$.

Algorithm 1 Replay with Stochastic Neural Transformation

REQUIRE: EEG decoder parameters $\boldsymbol{\theta}$, stochastic neural transformation parameters $\boldsymbol{\phi}$, EEG decoder learning rate η, neural transformation update rate β; transformation intensity G at each iteration, stored signal segments for replay $\mathcal{M}$;
for $i = 1$ to K **do**
  current signal segment $(\boldsymbol{x}_i, y_i)$ input to model.
  retrieve stored signal segment for replay, i.e., $(\boldsymbol{x}_m, y_m) \sim \mathcal{M}$
  perform stochastic neural transformation $\boldsymbol{x}_m(G) = \text{Transform}(\boldsymbol{x}_m, 0, G)$ (Eq. (3))
  update $\boldsymbol{\phi}$ with gradient ascent on $\max_{\boldsymbol{\phi}}[\mathcal{L}(\boldsymbol{x}_m(G), \boldsymbol{y}_m, \boldsymbol{\theta}, \boldsymbol{\phi}) - \mathcal{L}(\boldsymbol{x}_m, \boldsymbol{y}_m, \boldsymbol{\theta}, \boldsymbol{\phi}) - \lambda \mathbb{JS}(\boldsymbol{x}_m, \boldsymbol{x}_m(G))]$
  replay of $(\boldsymbol{x}_m(G), \boldsymbol{y}_m)$ and jointly train with $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for update of model parameters $\boldsymbol{\theta}$: $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \eta \nabla_{\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{\theta}_i, \boldsymbol{x}_m(G), y) + \mathcal{L}(\boldsymbol{\theta}_i, \boldsymbol{x}_i, y_i)]$
  update memory $\mathcal{M}$ by reservoir sampling (RS), $\mathcal{M} = \text{RS}(\mathcal{M}, (\boldsymbol{x}_i, y_i))$
end for

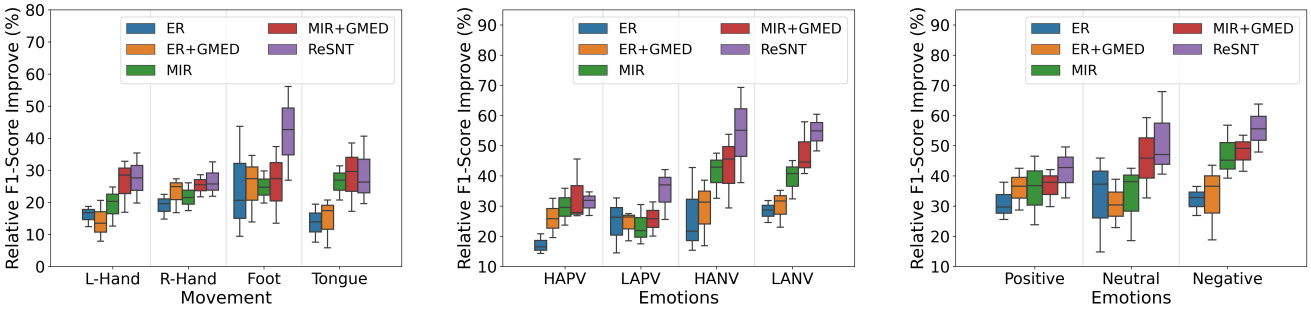
III. Empirical Evaluation

We constructed numerous sequential EEG decoding benchmarks for the problem setting on top of three large public datasets (BCI-IV 2a[16] [2], DEAP[17] [3] and SEED[18] [4]), which mimics the real world scenario of online sequential EEG decoding. Detailed ablation studies are performed in terms of transformation depth, memory size and different subject

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

TABLE I: Performance on accuracy and BWT evaluated after decoder finished sequential learning on all subjects.

Dataset	BCI-IV 2a		DEAP		SEED	
Method	Accuracy	BWT	Accuracy	BWT	Accuracy	BWT
EWC	41.18±0.75	−12.64±1.03	40.63±0.98	−11.41±0.57	46.53±1.74	−13.68±0.73
UCB	40.72±0.66	−13.14±0.58	38.24±1.31	−12.16±2.42	47.55±2.67	−12.94±2.35
ER	43.21±0.49	−12.56±0.72	41.45±1.80	−8.28±1.73	54.21±0.90	−10.05±1.16
ER+GMED	46.57±0.73	−12.10±0.51	43.19±2.03	−9.41±0.96	56.25±2.41	−11.98±0.85
MIR	48.35±1.14	−9.43±0.87	43.68±1.27	−7.55±2.53	58.64±1.56	−9.75±1.47
MIR+GMED	49.42±1.53	−7.49±1.96	44.52±0.74	−8.08±2.35	59.97±1.45	−8.36±2.33
ReSNT	**52.84±0.85**	**−6.13±1.24**	**46.20±1.96**	**−7.33±2.94**	**61.29±1.18**	**−6.95±1.34**
sequential	36.46±0.39	−14.78±0.41	33.41±0.92	−16.46±0.90	38.53±0.27	−18.84±0.91
joint train	81.70±0.52	—	71.57±0.69	—	82.50±0.86	—



(a) BCI-IV 2a (b) DEAP (c) SEED

Fig. 3: F1-score improvement for different classes, relative to the baseline of sequential learning directly on base decoder. The x axis corresponds to different classes of decoding tasks. For DEAP dataset, the high/low arousal are abbreviated as HA/LA, positive/negative valence are abbreviated as PV/NV, e.g. high arousal positive valence corresponds to "HAPV".

ordering, etc., which offers in-depth understanding of model performance.

Baselines We incorporated a wide range of baselines for comparison in the experiment, including: 1) upper and lower bound: lower bound is sequential learning across different subjects directly with base decoder, and the upper bound is joint training with data from all subjects simultaneously available. 2) currently widely used continual learning approaches, including both regularization approaches such as **EWC** [19], **UCB** [20], and replay-based approaches including **ER** [11], **MIR** [21] and **GMED** [22].

Benchmark and Metrics We created new benchmarks to mimic the scenario of continual decoding over a long period of time across sequentially arriving subjects. Three different scenarios of subject ordering is explored in the ablation study, including 1) sequential order based on subject id, which is the default setting in experiments, 2) ascending order based on decoding difficulty, and 3) descending order based on decoding difficulty. We fed the streaming EEG signal of different subjects sequentially into the decoder based on the aforementioned subject ordering.

We evaluated the average testing accuracy on all subjects after completion of sequential learning, and performed measurement on the improvement of F1 score per individual class. We also performed evaluation on backward transfer (BWT). BWT is defined as BWT $= \frac{1}{N-1}\sum_{i=1}^{N-1} a_{N,i} - a_{i,i}$, with $a_{j,i}$ being the accuracy evaluated on subject j after sequential learning of subject i. N is the total number of subjects. Negative value of BWT indicates the occurrence of catastrophic forgetting, and positive value shows learning of later subjects helps improve model performance on previous subjects.

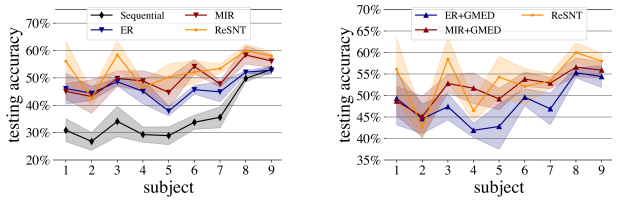
Model Settings We adopt the shallow ConvNet architecture introduced in [3] to serve as base model for the decoding task. We used a small memory to store 200 signal segments by default for both datasets. The diffusion and drift network for the neural transformation is formed with two layers of ConvNet with filter size 4×4 for first layer and 2×2 for second layer. We set the transformation intensity G to be 0.1 by default, with its influence on model performance studied in ablation study. The memory size is 200 for BCI-IV 2a and DEAP datasets, and we used a memory size of 500 for SEED dataset due to it is significantly larger. We set the regularization weight λ to be 2.0 by default, and its sensitivity is also examined in ablation study. Other hyperparameter settings follow [3]. The reported results are repeated for 10 runs in each setting.

A. Performance Evaluation

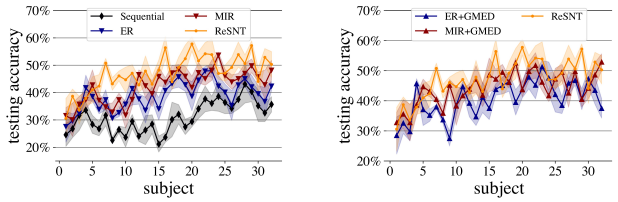
Table I shows the performance in terms of test accuracy and BWT on ReSNT and comparison baselines. The accuracy is evaluated on all subjects after sequential learning finishes. ReSNT outperforms the baselines by a significant margin on all three datasets. The gain on accuracy is 3.42%, 1.68% and 1.32% for BCI-IV 2a, DEAP and SEED respectively. It has a margin of 1.36%, 0.75% and 1.41% on the three datasets in terms of BWT. The model is better at maintaining

TABLE II: Ablation study on the influence of different subject ordering towards replay-based approaches. We explored three scenarios, 1) sequential order of subject id, 2) ascending order of decoding difficulty, 3) descending order of decoding difficulty.

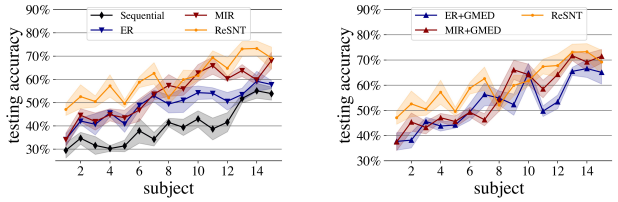
Scenario	ER	ER+GMED	MIR	MIR+GMED	ReSNT
BCI-IV 2a Dataset					
Sequential	43.21±0.49	46.57±0.73	48.35±1.14	49.42±1.53	52.84±0.85
Ascending	43.56±0.72	46.44±1.25	48.69±0.92	49.51±0.87	53.10±1.39
Descending	42.75±1.08	46.10±0.51	48.13±0.76	49.24±1.08	52.43±0.61
DEAP Dataset					
Sequential	41.45±1.80	43.19±2.03	43.68±1.27	44.52±0.74	46.20±1.96
Ascending	42.28±0.74	42.63±1.35	44.20±0.85	45.95±1.19	46.71±1.55
Descending	40.57±1.13	42.26±1.27	43.32±2.28	44.03±1.62	46.14±0.80
SEED Dataset					
Sequential	54.21±0.90	56.25±2.41	58.64±1.56	59.97±1.45	61.29±1.18
Ascending	55.53±1.42	57.71±1.25	58.89±0.73	59.85±1.28	61.84±1.63
Descending	53.64±1.75	55.91±1.03	58.28±1.39	59.36±0.93	60.97±1.10



(a) BCI-IV 2a Dataset



(b) DEAP Dataset



(c) SEED Dataset

Fig. 4: Model performance on individual subject after sequential learning ends. The proposed model maintains relatively good performance on earlier subjects with forget mitigation mechanisms. (a) BCI-IV 2a dataset, (b) DEAP dataset, (c) SEED dataset.

performance of previous subjects during sequential learning as it produces more robust and diversified data in feature space for effective replay. We performed further decomposition of model performance at class-level, in terms of the different motor imagery and emotion recognition tasks. The result is summarized in fig. 3. The relative improvement in F1 score for the different classes shows the proposed model has varying levels of effectiveness for the different recognition tasks. Specifically, the model shows it is more effective on capturing foot movement compared to other baselines

TABLE III: Effect of transformation intensity. We observed the performance tend to converge with transformation intensity $G > 0.1$.

Trans. Intensity	$G = 0.05$	$G = 0.1$	$G = 0.15$	$G = 0.2$
BCI-IV 2a	52.25±0.92	52.84±0.85	53.29±1.31	53.36±0.62
DEAP	45.57±0.63	46.20±1.96	46.41±1.22	46.44±1.59
SEED	61.04±1.35	61.29±1.18	61.67±0.80	61.75±1.37

when performing motor imagery decoding on BCI-IV 2a. For emotion recognition tasks, the model performs better than baselines in detecting high arousal negative valence emotions. We performed T-SNE visualization of transformed data, shown in fig. 5. We observed the transformed data is in general more scattered and shows a loosely mixed pattern in feature space. Fig. 4 shows the performance of individual subjects after the sequential decoding finished. The proposed ReSNT model is better at maintaining performance of earlier subjects, with more effective knowledge retaining during the sequential decoding process.

B. Ablation Study

Influence of Different Subject Ordering

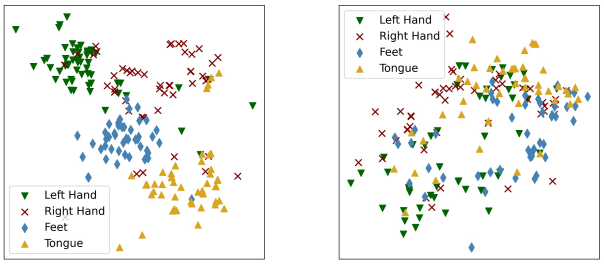
We explored the influence of different subject ordering on model performance. We run the model with three different subject ordering scenarios, ordered by 1) subject ID, 2) descending of decoding difficulty, 3) ascending of decoding difficulty respectively. The decoding difficulty is reflected as decoding accuracy during test. We summarized the result in Table II. Overall, the ascending order based on decoding difficulty turns out to yield the best overall performance, and descending order is the most challenging setting. This shows the model benefits from learning harder subjects at the beginning during the sequential decoding of different subjects.

Effect of Transformation Intensity

We performed evaluation of model performance with respect to different transformation intensities depicted by G. The result is provided in Table III. We observed the model performance improves monotonically with the increase of G, this shows the effectiveness of consistency regularization which helps preserve the similarity between transformed data and original data. The model performance is trending towards convergence with $G > 0.1$. We used $G = 0.1$ as default setting in our experiments.

Influence of Memory Size

We performed ablation study on different memory sizes for all replay based approaches. The result is shown in Table IV. The performance improves for the replay based approaches with larger number of signal segments stored in memory. We observed the ReSNT model has a larger margin compared to comparison baselines in small memory settings, e.g. for BCI-IV 2a dataset, ReSNT outperforms other baselines by at least 3.93% when only 100 segments are stored for replay, and this margin gradually shrinks to 0.57% when 500 segments are allowed in memory. This shows the ReSNT model suits better to scenarios with smaller memory settings. We used a memory size of 200 for BCI-IV 2a and DEAP datasets, and memory size of 500 for SEED dataset as default settings.



(a) Original Segments (b) Transformed Segments

Fig. 5: TSNE visualization at feature level of (a) original stored segments, (b) transformed segments in memory for the different motor imagery classes of BCI-IV 2a dataset. The stochastic neural transformation on stored segments produces more diverse and robust features for model generalization improvement.

TABLE IV: Model performance with respect to different number of signal segments in memory. The model shows improved avg. accuracy on all subjects with larger number of segments stored in memory.

Mem. Size	ER	ER+GMED	MIR	MIR+GMED	ReSNT
BCI-IV 2a Dataset					
100	40.63±1.32	42.49±0.85	45.96±2.01	46.22±0.79	50.15±1.46
200	43.21±0.49	46.57±0.73	48.35±1.14	49.42±1.53	52.84±0.65
500	51.47±0.80	53.28±1.79	56.82±1.53	57.56±0.85	58.13±0.92
DEAP Dataset					
100	35.80±0.76	37.64±0.95	39.56±2.89	40.10±3.24	42.63±1.70
200	41.45±1.80	43.19±2.03	43.68±1.27	44.52±0.74	46.20±1.96
500	48.64±0.59	50.42±0.39	52.26±2.16	53.97±1.91	55.15±1.34
SEED Dataset					
200	54.21±0.90	56.25±2.41	58.64±1.56	59.97±1.45	61.29±1.18
500	63.51±1.25	65.17±0.57	66.69±2.04	67.36±2.73	68.53±1.45
800	68.82±1.62	69.39±0.94	71.82±1.59	72.28±2.52	72.71±1.74

IV. Conclusion

For clinical applications of BCI systems, it is necessary for the system to perform decoding over a long period of time on streaming brain signals for patient assistance tasks. In this work, we propose a replay based approach for effective brain signal decoding in this challenging scenario. The model performs stochastic neural transformation on stored segments to generalize well on previous subjects and avoids overfitting. We formed a bi-level optimization framework for end to end training of the stochastic neural transformer together with the EEG decoder. Its effectiveness is evaluated on numerous newly formed benchmarks that mimic real world prolonged decoding scenarios. Further exploration is needed for the following items: 1) decoding of subject sequences with heterogeneous classes, 2) online continual decoding without usage of stored signal segments in memory.

V. Acknowledgements

This material is based upon work supported in part by the US National Science Foundation, Award #s: 1920920, 2125872 and 2223793.

References

[1] I. Iturrate, J. Antelis, and J. Minguez, "Synchronous eeg brain-actuated wheelchair with automated navigation," in *2009 IEEE International Conference on Robotics and Automation*, 2009, pp. 2318–2325.

[2] A. Campbell et al., "Neurophone: Brain-mobile phone interface using a wireless eeg headset," in *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems*, 2010.

[3] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[4] T. Duan, Z. Wang, S. Liu, Y. Yin, and S. N. Srihari, "Uncer: A framework for uncertainty estimation and reduction in neural decoding of eeg signals," *Neurocomputing*, vol. 538, 2023.

[5] 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," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, jul 2018.

[6] T. Duan and S. N. Srihari, "Layerwise interweaving convolutional LSTM," in *Advances in Artificial Intelligence - 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada*, vol. 10233, 2017, pp. 272–277.

[7] W.-L. Zheng and B.-L. Lu, "Personalizing eeg-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, pp. 2732–2738.

[8] T. Duan, M. A. Shaikh, M. Chauhan, J. Chu, R. K. Srihari, A. Pathak, and S. N. Srihari, "Meta learn on constrained transfer learning for low resource cross subject eeg classification," *IEEE Access*, vol. 8, pp. 224 791–224 802, 2020.

[9] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in Neural Information Processing Systems*, 2017.

[10] Z. Wang, T. Duan, L. Fang, Q. Suo, and M. Gao, "Meta learning on a sequence of imbalanced domains with difficulty awareness," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8927–8937, 2021.

[11] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, and M. Ranzato, "Continual learning with tiny episodic memories," *CoRR*, vol. abs/1902.10486, 2019.

[12] Z. Wang, L. Shen, T. Duan, D. Zhan, L. Fang, and M. Gao, "Learning to learn and remember super long multi-domain task sequence," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7972–7982, 2022.

[13] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *International Conference on Learning Representations*, 2018.

[14] R. Gontijo-Lopes, S. Smullin, E. D. Cubuk, and E. Dyer, "Tradeoffs in data augmentation: An empirical study," in *International Conference on Learning Representations*, 2021.

[15] P. Kidger, J. Foster, X. C. Li, and T. Lyons, "Efficient and accurate gradients for neural sdes," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 18 747–18 761.

[16] M. Tangermann et al., "Review of the bci competition iv," *Frontiers in Neuroscience*, vol. 6, p. 55, 2012.

[17] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis ;using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan 2012.

[18] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[19] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, 2017.

[20] S. Ebrahimi, M. Elhoseiny, T. Darrell, and M. Rohrbach, "Uncertainty-guided continual learning with bayesian neural networks," *arXiv preprint arXiv:1906.02425*, 2019.

[21] R. Aljundi, L. Caccia, E. Belilovsky, M. Caccia, M. Lin, L. Charlin, and T. Tuytelaars, *Online Continual Learning with Maximally Interfered Retrieval*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[22] X. Jin, A. Sadhu, J. Du, and X. Ren, "Gradient based memory editing for task-free continual learning," *ArXiv*, vol. abs/2006.15294, 2020.