Robust DNN-based Decoder Model with an Embdeded State-Space Model Layer

Pedram Rajaei
Department of Biomedical Engineering, University of Houston, TX, USA
prajaei@CougarNet.uh.edu

line 1: 2nd Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address

line 1: 3rd Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address

line 1: 4th Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address

line 1: 5th Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address

Ali Yousefi
Department of Biomedical Engineering, University of Houston, TX, USA
aliyousefi@uh.edu

***Abstract*— A critical step to probe brain function is characterizing time-series data, representing biobehavioral signals recorded under different conditions or cognitive tasks. The inherent complexity, volatility, and stochasticity present in these data pose significant modeling challenges. More recently, the deep neural network (DNN) models have demonstrated to be powerful tools in characterizing such a data, and thus are widely adopted in the neuroscience data analysis applications. However, DNNs have their own limitations as they show a higher sensitivity to noise present in data, and also require larger datasets for their training. These issues limit their benefits in many neuroscience data analysis problems as size of available data is limited, and recorded data show significant volatility and noise patterns. In this research we introduce novel framework which brings a state-space models (SSMs) as a part of DNN structure, which can address aspects of these limitations. We call this modeling solution SSM-DNN, and show attributes of this modeling solution which can overcome issues related to data size and noise. For the model, we develop its training and inference steps and show its application in a dataset recorded during an Implicit Association Test (IAT) task. The objective of this task is decoding participants phenotype, e.g. Major Depressive Disorder (MDD) versus Healthy. We show SMM-DNN performance reaches a specificity and sensitivity of XXX and YYY with an area under curve (AUC) of ZZ, surpassing state-of-art DNN models. SSM-DNN introduce a powerful decoder model, which can be applied to a broader class of neuroscience data.**

Keywords*— *Neural Decoder, Bayesian Inference, Adaptive Dimensionality Reduction, Supervised State Space Model, DNN

I. Introduction

Deep neural network models and its variants such as CNN and RNN have revolutionized many domains of science, achieving state-of-the-art performance in many predictive and classification tasks [1]. In neuroscience, we can point to many applications of DNN models such decoding EEG and fMRI data to probe cognitive states or categories present in the data [2][3]. However, DNNs superb performance is highly dependent on large-scale and high-quality labeled dataset, requiring significant resources and time. Neuroscience data are generally limited in size, and come with different source of noise and volatility, which make reaching a high level of accuracy and robustness of decoding using DNN a challenge. Potential solutions here include data augmentation [4], advanced pre-processing techniques [5], and the use of generative models.

In this research, we aim to address the aforementioned challenges by using state-space model (SSM) as a part of DNN structure. SSMs have shown to be a powerful tool for adaptive noise reduction, and it is generally used to infer a dynamical low dimensional representation of the data, capturing essential dynamics present in data. SSMs are a generative model; thus, we can use to create proper augmentation data. Instance of these aspects are shown in previous work such as Youefi et al [] and XXX & YYY [ref].

We propose a hybrid framework that integrates SSM and DNN in a unified modeling framework. We show SSM-DMM inherits aforementioned ccapabilities of SSMs and combines it with expressiveness and discriminative attributes of DNNs. We define the model structure, and develop its training and inference of SSM-DNN. We also demonstrate its application in a dataset, and show the model capability to address DNN challenges and while attaining a superb decoding accuracy.

In the following sections, we first introduce SSM-DNN followed by its training and inference step. We then introduce IAT task, and then how our modeling results along with

This project is sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No. N660012324016. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

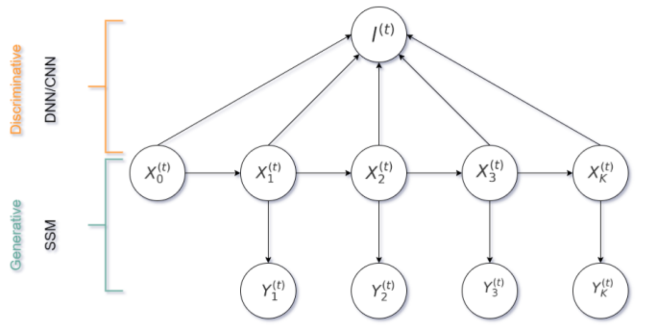

Fig. 1. SMM-DNN Model Architecture The model combines SSM and DNN in characterization of high-dimensional neural recording and their labels. x_k^t represent a low dimensional representation of neural data, which is passed to a DNN for supervised learning tasks, enabling accurate label prediction and interpretation. This integration leverages the generative strengths of SSMs for capturing temporal dynamics while utilizing the discriminative capabilities of DNNs for classification and regression.

performance of a DNN in the predicting MDD and CTL given the behavioral data.

II. Methods

A. Model Definition

Let us consider an experiment conducted over T trials (t=1, …, T). During each trial, the observed data consist of biobehavioral signals, Y^t={$y^t_{1\ldots K}$} which are recorded at K time points and $Y_k^t \in R^n$, and label l^t that represents one of C classes, defined by $l^t \in \{1, \ldots, C\}$, representing the experiment condition or stimulus. We assume the distribution of observed data and label are defined as follow:

$$y_k^t | x_k^t \sim g_\phi(x_k^t) \quad (1.a)$$

$$l^t | x_{0\ldots k}^t \sim h_\tau(x_{0\ldots k}^t) \quad (1.b)$$

Specific case of the observed data conditional distribution can be defined by $y_k^t | x_k^t \sim N(g_\phi(x_k^t), Q)$, where we assume this distribution follows a multivariate normal distribution with an additive Gaussian noise.

In equation set (1), x_k^t is a latent variable, representing the underlying dynamical manifold shaped both observed data and label. The evolution of the latent variable over time is defined by a state equation:

$$x_{k+1}^t | x_k^t \sim f_\psi(x_k^t) \quad (2)$$

Specific case of state process equation is $x_{k+1}^t | x_k^t \sim N(f_\psi(x_k^t), R)$, where we assume this conditional distribution follows a multivariate normal distribution with an additive Gaussian noise. The latent variable has dimension D, $x_k^t \in R^D$, which are generally much smaller than N.

Equation set 1 & 2 define SSM-DNN model. The model parameters are ω ={ϕ, τ, ψ}, which we will discuss in the next section how it can be estimated. **Fig. 1** depicts the model structure. The latent process changes from one time point to the next, and the biobehavioral signals evolve at the same time scale as the latent process. The whole trajectory of the latent process enters into the DNN model to create the label or category. When the model parameters W, are known, we can leverage from generative nature of the model to create data.

B. Model Training

The model training corresponds to maximizing the evidence of both label and osbserved data over all trials of the experiment. We assume each trial is independent from the other one; thus, the evidence is defined by

$$P_\omega(Y, L) = \int_{X^T} \ldots \int_{X^0} \prod_{t=1}^{T} P_\omega(Y^t, l^t; X^t)\, dX^0 \ldots dX^T \quad (3)$$

where, $Y = \{Y^{1,\ldots,T}\}$, $L = \{l^{1,\ldots,T}\}$ and $X^T = \{X^T_{0,\ldots K}\}$.

Given the evidence, the model training corresponds to finding a parameters set that maximizing it. This is defined by

$$\omega = argmax P_\omega(Y, L) \quad (4)$$

To find the parameters, we first need to take the integral over latent variable – Eq. (3), which is computationally expensive and it becomes an interactable problem. In this case, we can use Expectation-Maximization (EM)[11] or Variational Inference (VI) approaches to detour this integration and estimate the model parameters [12]. Here, we used EM for the model training, which turns into two modeling steps: expectation (E-step) and maximization (M-step). For these two steps, we require the posterior distribution of latent processes given the observe data and label. We use MCMC approach to approximate this posterior, and then use inferred trajectories of the posterior (particles in time) to approximate expectation of the logarithm of full likelihood, E-step [13]. **Table 1** describes the MCMC approach for SSM-DNN. With the E-M, the model training turns into an iterative procedure switching between E- and M-steps. For the M-step, we use stochastic gradient ascent to update the model parameters. **Table 2** summarizes the training process. Details of implementation can be found in GitHub at: https://github.com/Pedram-Rajaei/LDCM. Details of implementation can be found in GitHub at: https://github.com/Pedram-Rajaei/LDCM

C. Decoding Process

We define the decoding process as predicting probability of the label per trial given the model parameters and observed biobehavioral data. The process starts by finding the posterior distribution of state given the observed data – $P(X^t \mid Y^t)$, and we then predict probability of different classes given multiple trajectories of the inferred state. In contrast with the training process, the state inference is defined given the biobehavioral signals and without label. We argue that the model parameter, updated in the training step, will push the state and observation parameters into a domain that lets the state inference to be amenable to the category or label. Simply, the inference plays the rule of an adaptive feature extraction step.

In the next section, we demonstrate an application of our framework on IAT dataset captured in our research group at University of Minnesota to probe behavioral deviation in Major Depressive Disorder (MDD) participant group.

Table 1 MCMC approach for SSM-DNN

- **Initialization**: Select an initial estimate for the unknown parameters, $\Theta^{(0)}$

- **E-step:** Using the point process filter and smoother, compute the conditional expectation of the log likelihood function

$$l(\Theta|\Theta^{(m)}) = \mathrm{E}_{p(\boldsymbol{x}_{0\cdots N}|\Delta N_{1\cdots N};\Theta^{(m)})}[\log L(\boldsymbol{x}_{0\cdots N}, \Delta N_{1\cdots N}; \Theta)]$$

where, $L(\boldsymbol{x}_{0\cdots N}, \Delta N_{1\cdots N}; \Theta)$ is the full-likelihood function and $p(\boldsymbol{x}_{0\cdots N}|\Delta N_{1\cdots N}; \Theta^{(m)})$ comes from the smoother distribution of $\boldsymbol{x}_{0\cdots N}$ given the spiking observations and previous parameter estimates.

-**M-step:** Find a new set of parameters that maximizes the $l(\Theta|\Theta^{(m)})$

$$\Theta^{(m+1)} = \underset{\Theta}{\operatorname{argmax}}\, l(\Theta|\Theta^{(m)})$$

-**Neural network step:** Update Neural Network Parameters

Table 2 SSM-DNN Training process

- **Initialization:** For $u = 1 \cdots U$ particles, draw samples of the initial state $\boldsymbol{x}_0^{(u)}$ from the initial density $p(\boldsymbol{x}_0)$, and set $\tilde{w}_0^{(u)} = U^{-1}$ for all u.

- **Importance sampling step:** At each time step, for $u = 1 \cdots U$, sample $\hat{\boldsymbol{x}}_k^{(u)}$

from $\pi_k(\boldsymbol{x}_k; \boldsymbol{x}_{k-1}^{(u)}, \Delta N_{1\cdots k})$ and set $\hat{\boldsymbol{x}}_{0:k}^{(u)} = (\boldsymbol{x}_{0:k-1}^{(u)}, \hat{\boldsymbol{x}}_k^{(u)})$. Compute the importance

weights

$$w_k^{(u)} = \tilde{w}_{k-1}^{(u)} \frac{p(\Delta N_k|\hat{x}_k^{(u)}, \Delta N_{1\cdots k-1})\, p(\hat{x}_k^{(u)}|x_{k-1}^{(u)})}{\pi_k(\hat{x}_k^{(u)}; x_{k-1}^{(u)}, \Delta N_{1\cdots k})}$$

and normalize the importance weights

$$\hat{w}_k^{(u)} = w_k^{(u)} \left[\sum_{u=1}^{U} w_k^{(u)}\right]^{-1}$$

- **Resampling step:** Resample with replacement U particles $\boldsymbol{x}_{0:k}^{(u)}, u = 1 \cdots U$ from the set $\left(\hat{\boldsymbol{x}}_{0:k}^{(u)}, u = 1 \cdots U\right)$ with probabilities determined by the normalized importance weights $\hat{w}_k^{(u)}$ [47], [59]. Reset the weights to $\tilde{w}_k^{(u)} = U^{-1}$.

III. Data Description

The dataset used in this study comprises behavioral data collected during an Implicit Association Task (IAT) designed to investigate potential behavioral and cognitive differences between individuals with MDD and healthy controls (CTL). The data collection was conducted under IRB approval (Advarra IRB, approved on 01/05/2024). A total of 23 participants data out of XX data was used in this research (remaining data were excluded as participants missed more than XX trials and their performance accuracy was below chance level), including 11 individuals diagnosed with MDD and 12 healthy controls without any reported psychiatric conditions.

The experimental task required participants to match a stimulus word with one of two categories displayed as the header on the screen: "Life + Me" or "Death + Me.". Stimulus words were sequentially presented on the screen, and participants were instructed to press the corresponding button to indicate whether the word matched or did not match the active header. Participants had 2.5 seconds to respond, and the task design alternated the header categories every 20 trials to reduce potential response bias and introduce variability. The stimulus words were carefully selected to reflect semantic associations with the task categories Words such as: I, Myself, Alive, Happy, etc. were chosen to be corresponded by Life + Me and Words such as: Die, Dead, Lonely, Sad, etc. were chosen to be corresponded by Death + Me.

Each participant completed 360 trials, organized into 18 blocks, with each block containing 20 trials. Reaction times (RT) and accuracy were recorded for each trial. In this research we use RT data to examine in what extent participant phenotype, MDD vs CTL, are encoded in the RT data.

IV. MDD vs CTL Decoder Model

For the decoder model, the observed data is RT time series with the length of 360. and the label is either of MDD or CTL categories. We assume the latent process is a random walk model, defined by

$$x_{k+1} = x_k + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma_v^2) \tag{5}$$

The observation and label model is defined by

$$y_k = x_k + v_k, \quad v_k \sim \mathcal{N}(0, Q) \tag{6}$$

$$l^t | x_{0,\ldots,K}^t \sim 1D - CNN(\rho) \tag{7}$$

Here, we dropped superscript of t, referring to a trial from the test session or pool. The DNN utilized is one-dimensional CNN with ρ as its hyper parameters. XXX.

For this specific problem, we use a simple linear state equation model. As such, we expect SSM will mainly contribute to regressing the noise for the observation and path a smoothed trajectory of RT to the DNN model. However, the model definition and its inference process can scale properly to higher dimension of the observation and more complex state equation. We will discuss this aspect in another research and it is not the main scope of this research. noise. However,.

V. Performance Results

To assess the model decoding accuracy, we partitioned dataset into training and testing sets. Specifically, we allocated 6 participants (3 MDD and 3 CTRL) for the test set, while the remaining 16 participants (8 MDD and 8 CTRL) were used for training. We drop one participant data to create a balanced number of each group sample in the training, making sure the model training and test are unbiased. We create these partitions randomly and repeat 10 times. We then assess the performance by averaging decoding accuracy in the test sets.

To evaluate the model's performance, we plotted the Receiver Operating Characteristic (ROC) curve. The results demonstrated a significant improvement in classification performance when compared to the same model trained using a transfer learning approach. The ROC curve generated from our primary model achieved a much higher area under the curve (AUC), indicating superior sensitivity and specificity in distinguishing between MDD and CNTR participants. These findings suggest that our tailored model, trained specifically on this dataset, is better suited for the given task compared to a transfer learning approach, which likely suffered from domain mismatch and overfitting to unrelated features.

VI. Discussion

In this research, we introduce a novel decoder framework which are called SSM-DNN. SSM-DNN address challenges of the DNN which includes sensitivity to noise and requiring large dataset. Our proposed solution reaches XXXX performance which is significantly higher than YYYYY. We argue the proposed solution can be expanded to multivariate time-series such as EEG with multiple channels of recording, where the laten process finds a low dimensional manifold, or neural features, amiable to the condition or label of the data. This is a critical modeling step as the dimensional reduction or feature extraction is done in tandem with the DNN training instead of running disjoint steps. Besides, the inference process is not a black box and gives a reasonable amount of interpretability about the data which is critical in neuroscience experience and decoding processes.

VII. Future Work

While EEG signals were recorded from multiple channels, embedding EEG signal in the decoding pipeline was not scope of this research. For the next, we aim to expand the latent process to combine both behavior and neural date, EEG here, in predicting participants' labels.

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436–444

[2] Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering, 15(5), 056013.

[3] Chaudhary, S., Pachori, R. B., & Upadhyay, A. (2019). Deep learning-based classification of fMRI data for mental state decoding. IEEE Transactions on Cognitive and Developmental Systems, 11(1), 46-55.

[4] Ullah, H., Anwar, S. M., & Bilal, M. (2021). Data Augmentation for Improving Performance of Deep Neural Networks in EEG-Based Emotion Recognition. Frontiers in Computational Neuroscience, 15, 723843.

[5] Zhao, X., Zhang, X., & Zou, J. (2021). Pre-processing Methods for Improving Robustness of EEG Classification in Deep Learning Applications. Frontiers in Human Neuroscience, 15, 765525.

[6] Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. NeuroImage, 58(2), 312-322.

[7] Moran, R. J., Kiebel, S. J., Stephan, K. E., Reilly, R. B., Daunizeau, J., & Friston, K. J. (2007). A neural mass model of spectral responses in electrophysiology. NeuroImage, 37(3), 706-720 .

[8] Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. The Lancet, 395(10225), 689-697.

[9] Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., & Paninski, L. (2019). Bayesian learning and inference in recurrent switching linear dynamical systems. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 914-922).

[10] Särkkä, S. (2013). Bayesian Filtering and Smoothing. Cambridge University Press. https://doi.org/10.1017/CBO9781139344203.

[11] Smith, A. C., & Brown, E. N. (2003). Estimating a state-space model from point process observations. Neural Computation, 15(5), 965-991.

[12] Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. Journal of the American Statistical Association, 112(518), 859-877.

[13] Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1995). Markov Chain Monte Carlo in Practice. Chapman and Hall/CRC.