2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Variational Encoder Framework for Decoding Behavior Choices from Neural Data

Shiva Salsabilian and Laleh Najafizadeh

Abstract— **In this paper, using an adversarial variational encoder model, we propose a two-step data-driven approach to extract cross-subject feature representations from neural activity in order to decode subjects' behavior choices. First, various characteristics of the recorded behavior are computed and passed as features to a clustering model in order to categorize different behavior choices in each trial and create labels for the data. Then, we utilize a variational encoder to learn the latent space mappings from neural activity. An attached adversary network is used in a discriminative setting to detach the subject's individuality from the representations. Recorded cortical activity from Thy1-GCaMP6s transgenic mice during a motivational licking experiment was used in this study. Experimental results demonstrate the capabilities of the proposed method in extracting discriminative representations from neural data to decode behavior by achieving an average classification accuracy of 88.8% across subjects.**

I. INTRODUCTION

Understanding the relationship between neural activity and behavior has been a challenging neuroscience research problem. To approach this problem, conditioned and simplistic paradigms have been typically used to make evaluating the relationship between the behavior (e.g. decision, memory, learning) and the brain function, feasible. Behavior is traditionally characterized with low-dimensional task-related variables. Once low dimensional features of the behavior are extracted, through a supervised or unsupervised method, dependency of the neural activity to these behavioral signals can be modeled. Classic approaches usually use simple quantitative measures that are easy to relate back to the experiment [1]. Recent approaches have begun to use machine learning and unsupervised methods to decompose more detailed behavioral measures [2]–[5].

A major challenge in neural decoding is the undesirable subject variability, which imposes difficulties in identifying features in neural signals that are common across subjects for decoding behavior. In general, to overcome the challenge of subject variability, promising results have been recently demonstrated using transfer learning approaches via autoencoders models [6]–[11]. For example, the cross-subject transfer learning approach in [10] aims to discover and exploit shared features that are invariant and generalizable across subjects. These methods rely on learning generative models of the data utilizing variational autoencoders (VAEs) that allow for the synthesis of data samples from latent representations for unsupervised feature learning, or generative adversarial networks (GANs) [12]. In another example, in [11], using adversarial learning within a deep encoder network, session-invariant person-discriminative representations were learned from EEG data for brain computer interfaces (BCIs).

In this study, we propose a two-step data-driven approach to decode behavior from neural data. Our approach is based on a deep autoencoder model, which includes the following steps. As the first step, we extract various characteristics of recorded behavior choices as feature vectors. Using a clustering algorithm, we categorize behavior choices in each trial and create labels for the data. Next, we utilize a variational encoder with an attached adversarial network approach for transfer learning to capture discriminative properties that are robust to subject variability for decoding behavior choices. The adversarial network aims to learn features from cortical activity timeseries, while the adversary network is used in a discriminative setting to detach the subject's individuality from the representations.

The remaining of this paper is organized as follows. Experimental procedures are described in Section II. Data analysis and methods are discussed in Section III. Experimental results are presented in Section IV, and the paper is concluded in Section V.

II. EXPERIMENTAL DATA

Widefield calcium imaging of fluorescence indicators in mice have been increasingly used in neursocience research, and has been utilized to study the relationship between the brain function and injury [13], [14], [15], [16] [17], as well as the relationship between the brain function and behavior [5], [18], [19].

The data used in this study was obtaiend from 8 subjects and was provided by the Department of Cell Biology and Neuroscience. The experiments were approved by the Rutgers University Institutional Animal Care and Use Committee. The data acquisition procedures are fully described in our previous work [5], [18], [19]. Briefly, cortical Ca^{2+} transient activity were recorded from Thy1-GCaMP6s transgenic mice via a customized microscope that provided the visualization of nearly the entire left hemisphere. Filtered fluorescence emission from the cortex was acquired using a MiCam Ultima CMOS camera at 100 frames per sec. Recordings were obtained from each mouse in two sessions. Each session, included 100 trials. Each trial consisted of a 0.9 s baseline followed by 1 s rising tone auditory stimulation. Water was

This work was supported by NSF award CBET-1605646.

Authors are with Integrated Systems and NeuroImaging Laboratory, Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA. {shiva.sal and laleh.najafizadeh}@rutgers.edu.

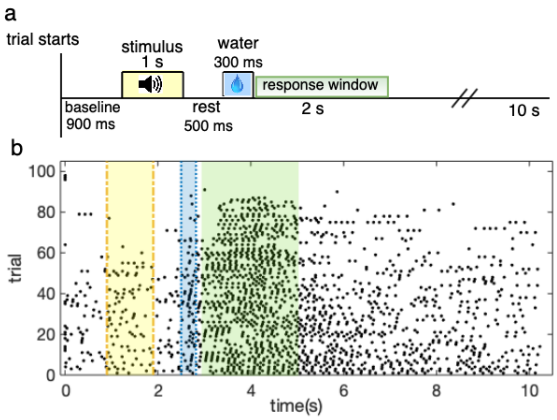

Fig. 1: (a) Timeline in each trial. (b) Dots indicate recorded lick responses. The duration of tone stimulus, water delivery and response window are shown in yellow, blue and green, respectively.

delivered 500 ms after the stimulus for the duration of 300 ms, and the trial was ended at 10.23 s (Fig. 1-a). Licking instances were recorded via a capacitive sensor.

III. Data Analysis and Methods

A. Extracting Cortical Activity Timeseries

Thirty 5×5-pixel regions of interest (ROIs)/channels ($C = 30$), distributed over the cortex, were selected based on their location relative to the bregma point. Timeseries associated with each channel were obtained by averaging the intensities of pixels within the 5×5-pixel regions, in each frame.

B. Behavior Clustering and Label Generation

As the session progresses, mice get more familiarized to the auditory stimulation and learn better that the time to lick is after the tone is played. An example of licking behavior from one subject in one session is shown in the Fig. 1-b. It can be observed that, as trials progress, the mouse appears to be more engaged in the experiment and shows signs of learning since the majority of the licks are occurring within the response window rather than random instances during the course of the trial. Similar behavior was observed for other mice, demonstrating that subject's lick behavior changes throughout the course of a session showing different stages of motivation, engagement, and learning.

To quantify the behavior of subjects within each trial and distinguish between the two possibilities —that is, motivated licking choices conditioned to the auditory stimulation as opposed to spontaneous licking —we used "lick rate" as the behavior/learning measure and used it to sub categorize the trials. For each trial, we calculated the lick rate for 4 segments of the trial as number of licks/duration of the segment. These 4 segments are the baseline (900 ms), the tone stimulus (1 s), the response window (2 s), and during the remaining of the trial.

A feature vector of $f = [f_1, f_2, f_3, f_4] \in \mathbb{R}^{1\times4}$ was generated for each trial. A series of *k*-means-based clustering algorithm with respect to the number of desired clusters was then built, and the best one with respect to the Silhouette value was identified. The best optimal Silhouette value was achieved with $k = 2$. The label of clusters formed by the k-means algorithm with $k = 2$, were used as the label for timeseries data of each trial.

C. Data Organization

For each trial, a sliding window with duration of T and step size of w time points was moved over the timeseries within the response window (2 s after water delivery). Data within each window was formed as a data matrix $\mathbf{X} \in \mathbb{R}^{C\times T}$. Gathering data from all the trials, the dataset $\{(\mathbf{X}_i^s, y_i^s)\}_{i=1}^{n_s}$, consisting of a total of n_s data samples, was obtained for subject s, where, $y_i^s \in \{0, 1\}$ represents the class label of behavior data generated for each trial. The attribute $s \in \{1, \ldots, S\}$ denotes the subject label among $\boldsymbol{s}$ subjects (i.e. a vector of size $1 \times S$ with one value 1 at s^{th} index, and 0 in other indices).

D. Model Architecture

Given the data, the aim is to build a discriminative decoder model that predicts y from observation data $\mathbf{X}$. To make the model generalizable across subjects, the prediction should be invariant to the attribute $\boldsymbol{s}$. We aim to enforce the latent representation to include minimum subject-dependent information via using an adversary network. This makes the model capable of achieving discriminative properties that are robust to subject variability and correspond to the common structure of the data shared among the subjects.

The network architecture involved variational autoencoder as an stochastic network, to minimize the reconstruction loss of the input data $\mathbf{X}$ and $\hat{\mathbf{X}}$. The encoder network is trained to learn representation $\boldsymbol{z} = g(\mathbf{X}, \theta_e)$. Obtained representations are used as inputs to a classifier with parameter θ_c to estimate y, and also as inputs to an adversary network with parameter θ_a, which aims to recover the variable s. The adversary network is trained to predict s by maximizing the likelihood $q_{\theta_a}(\boldsymbol{s}|\boldsymbol{z})$. At the same time, the encoder is trying to conceal the embedded information regarding $\boldsymbol{s}$ in the representation $\boldsymbol{z}$, as well as including sufficient discriminating information for the classifier to estimate y. These are achieved by minimizing the likelihood $q_{\theta_a}(\boldsymbol{s}|\boldsymbol{z})$ for the former and maximizing the likelihood $q_{\theta_c}(y|\boldsymbol{z})$ for the later objective. Therefore, the final objective function to train the proposed model structure simultaneously can be written as

$$arg \min_{\theta_c,\theta_e} \max_{\theta_a} \mathcal{L}(\theta_e, \theta_c, \theta_a) \quad (1)$$

$$\mathcal{L} = \mathbb{E}_z\mathbb{E}_y\ [-log\ q_{\theta_c}(y|\boldsymbol{z})] + \lambda\ \mathbb{E}_z\mathbb{E}_s\ [log\ q_{\theta_a}(s|\boldsymbol{z})], \quad (2)$$

where $\lambda \geq 0$ denotes the weight parameter that adjusts the impact of the adversary network. The optimization algorithm uses stochastic gradient descent for the adversary and the encoder-classifier networks to optimize (1) based on [10].

The structure of the proposed method is illustrated in Fig. 2 and is detailed in Table I. In our implementation, the temporal and spatial convolutional architectures are utilized

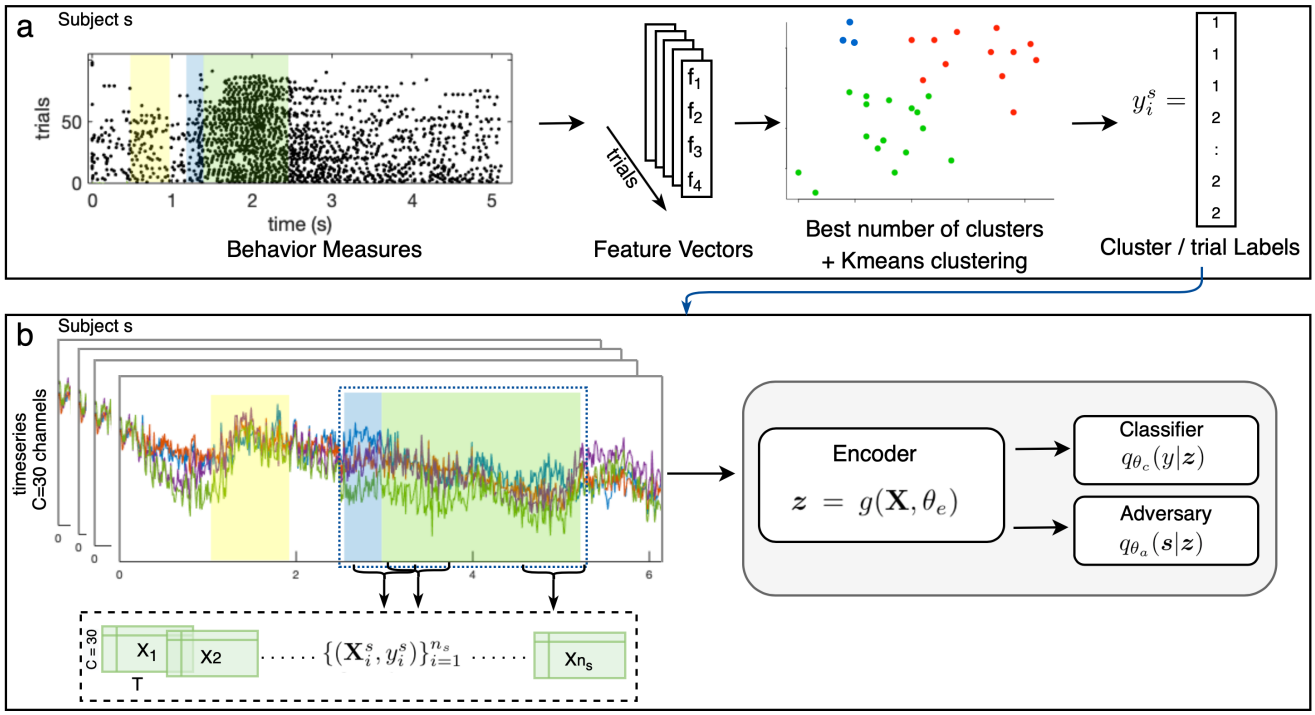

Fig. 2: (a) Schematic representation of the proposed autoencoder architecture for behavior feature extraction and labeling. (b) The proposed network model for behavior prediction based on cortical activity timeseries.

in the encoder architecture, embedding the temporal and spatial filtering. We used 4 temporal and 8 spatial convolutional units. The last fully-connected layer at the output of the encoder generates d_z-dimensional latent parameter vector. The classifier utilizes representation z as an input to a fully-connected layer with a softmax unit for class label discrimination. The adversary network is realized as a fully-connected layer with S softmax units for subject discrimination, to obtain normalized log-probabilities that will be used to calculate the losses. We used $W = 100$ for the temporal convolution kernel size, $C = 30$ for the spatial convolution kernel size, and the latent vector dimension of $d_z = 8$.

TABLE I: Encoder-Decoder Network Architecture

Encoder	$4 \times$ 1D Temporal Conv. $(1 \times W)$ + ReLU $8 \times$ 1D Spatial Conv. $(C \times 1)$ + ReLU (Reshape)+ Fully Connected layer $(20T \times d_z)$
Classifier	Fully Connected layer + softmax
Adversary	Fully connected layer + softmax $(d_z \times S)$

IV. Results

As discussed in Section III-B, different stages of behavior were noticeable during the course of the experiment. We identified these changes to discriminate the subject's behavior between the two possibilities of motivated licking choices conditioned to the auditory stimulus and spontaneous licking. We considered the licking rate in 4 different duration of the experiment (baseline, tone stimulus, response window, and remaining of the trial) as behavioral features. Fig. 3 shows the lick rate for these averaged over the trials from all subjects, labeled as motivated licking or spontaneous licking. We can observe that the lick rate at the baseline, during the tone stimulus, and the remaining of the trial considerably decreases in the motivated trials compared to the spontaneous licking trials. However, the lick rate during response window remains the same in both behavior categories.

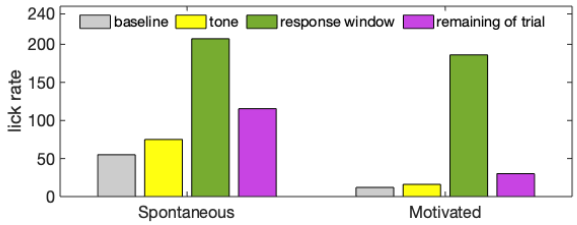

Fig. 3: Average lick rate at 4 different duration of the experiment over the trials of the two behavior categories.

The dataset $\{(\mathbf{X}_i^s, y_i^s)\}_{i=1}^{n_s}$ was collected by selecting the timeseries with duration $T = 400$ time points and step size of $w = 200$ and gathering the data matrices $\mathbf{X}_i \in \mathbb{R}^{30\times400}$ as shown in Fig. 2-b. Each $\mathbf{X}_i$ was labeled as motivated or spontaneous based on their corresponding category described in Section III-B. Considering the two recording sessions each with 100 trials, the dataset $\{(\mathbf{X}_i, y_i)\}_{i=1}^{n_s}$ of size $n_s = 1800$ is collected for each subject $s \in \{1, \ldots, 8\}$.

We randomly selected 2 subjects to hold-out for later cross-subjects transfer learning evaluation. For each analysis, we repeated this procedure 20 times and presented the averaged outputs. The network was trained using the remaining

subjects in a training-validation analysis with 80% for training and 20% for validation. Training data were normalized to have zero mean. The selected value ranges for the adversarial weight are $\lambda \in \{0, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. Note that these parameter combinations can be further optimized by cross-validating the model learning process.

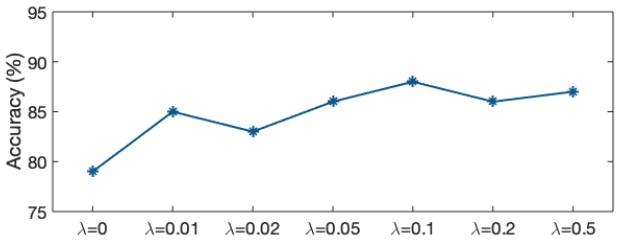

Fig. 4: Transfer learning average classification accuracy results for the 2 held-out subjects.

Cross-subjects analysis was performed to evaluate the trained model's transfer learning performance. Fig. 4 demonstrates the average transfer learning results for the 2 held-out subjects with different choices of λ. We observe that using the adversary network, ($\lambda \neq 0$), the accuracy result significantly improves, which emphasizes the added impact of the adversary network in achieving a more stable performance to decode data of unknown subjects by eliminating the subject dependent information from the representations. Note that $\lambda = 0$ eliminates the input of adversary network in the model training. The highest accuracy of 88.8% was achieved with $\lambda = 0.1$, demonstrating acceptable performance in cross-subject transfer learning of behavior decoding.

V. Conclusion

In this study, we proposed a framework based on autoencoder model for transfer learning in decoding behavior from neural data. We employed a two-step data-driven approach. First, various characteristics of recorded behavior were extracted as feature vectors and fed as input to a clustering model to create labels for the data. Next, we utilized a variational encoder with an attached adversarial network for cross-subject transfer learning which captures discriminative properties that are robust to subject variability. Experimental results demonstrated the benefits of the proposed framework in extracting robust subject invariant features and the capability of the proposed method for decoding behavior.

VI. Acknowledgment

We thank Professor David. J. Margolis and Dr. Christian. R. Lee, with the Department of Cell Biology and Neuroscience at Rutgers university, for acquiring and sharing the data.

References

[1] F. A. Wichmann and N. J. Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[2] S. R. Egnor and K. Branson, "Computational analysis of behavior," *Annual Review of Neuroscience*, vol. 39, pp. 217–236, 2016.

[3] A. B. Wiltschko, M. J. Johnson, G. Iurilli, R. E. Peterson, J. M. Katon, S. L. Pashkovski, V. E. Abraira, R. P. Adams, and S. R. Datta, "Mapping sub-second structure in mouse behavior," *Neuron*, vol. 88, no. 6, pp. 1121–1135, 2015.

[4] K. Raeisi, M. Mohebbi, M. Khazaei, M. Seraji, and A. Yoonessi, "Phase-synchrony evaluation of EEG signals for multiple sclerosis diagnosis based on bivariate empirical mode decomposition during a visual task," *Computers in Biology and Medicine*, vol. 117, p. 103596, 2020.

[5] L. Zhu, C. R. Lee, D. J. Margolis, and L. Najafizadeh, "Decoding cortical brain states from widefield calcium imaging data using visibility graph," *Biomedical Optics Express*, vol. 9, no. 7, pp. 3017–3036, 2018.

[6] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[7] S. Salsabilian and L. Najafizadeh, "An Adversarial Variational Autoencoder Approach Toward Transfer Learning for mTBI Identification," in *10th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2021.

[8] C. Tan, F. Sun, B. Fang, T. Kong, and W. Zhang, "Autoencoder-based transfer learning in brain–computer interface for rehabilitation robot," *International Journal Of Advanced Robotic Systems*, vol. 16, no. 2, p. 1729881419840860, 2019.

[9] H. Li, N. A. Parikh, and L. He, "A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes," *Frontiers in neuroscience*, vol. 12, p. 491, 2018.

[10] G. Louppe, M. Kagan, and K. Cranmer, "Learning to pivot with adversarial networks," *arXiv preprint arXiv:1611.01046*, 2016.

[11] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Adversarial deep learning in eeg biometrics," *IEEE signal processing letters*, vol. 26, no. 5, pp. 710–714, 2019.

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] F. Koochaki, F. Shamsi, and L. Najafizadeh, "Detecting mTBI by Learning Spatio-temporal Characteristics of Widefield Calcium Imaging Data Using Deep Learning," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 2917–2920.

[14] S. Salsabilian and L. Najafizadeh, "Detection of mild traumatic brain injury via topological graph embedding and 2D convolutional neural networks," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 3715–3718.

[15] S. Salsabilian, E. Bibineyshvili, D. J. Margolis, and L. Najafizadeh, "Study of functional network topology alterations after injury via embedding methods," in *Optics and the Brain*. Optical Society of America, 2020, pp. BW4C–3.

[16] ——, "Quantifying changes in brain function following injury via network measures," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 5217–5220.

[17] K. Tripathy, Z. E. Markow, A. K. Fishell, A. Sherafati, T. M. Burns-Yocum, M. L. Schroeder, A. M. Svoboda, A. T. Eggebrecht, M. A. Anastasio, B. L. Schlaggar, *et al.*, "Decoding visual information from high-density diffuse optical tomography neuroimaging data," *NeuroImage*, p. 117516, 2020.

[18] S. Salsabilian, L. Zhu, C. R. Lee, D. J. Margolis, and L. Najafizadeh, "Identifying task-related brain functional states via cortical networks," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–4.

[19] S. Salsabilian, C. R. Lee, D. J. Margolis, and L. Najafizadeh, "Using connectivity to infer behavior from cortical activity recorded through widefield transcranial imaging," in *Optics and the Brain*. Optical Society of America, 2018, pp. BTu2C–4.