TRUSTWORTHY DISEASE DISCOVERY: A CASE STUDY IN ALZHEIMER'S

Anonymous authors
Paper under double-blind review

ABSTRACT

A small number of high-dimensional training samples is a challenging problem for understanding the dynamics of multivariate spatiotemporal neuroimaging. Often, reducing these dynamics to a small number of handcrafted features helps. For example, the matrix of Pearson's correlation coefficients is highly predictive. Nevertheless, it is hard to perceive the dynamics and the disorder from these compressed proxy representations. In this paper, we propose a hierarchical recurrent model with attention that learns dynamics directly from temporal signals and captures stable interpretations of abnormal conditions predictive of disorder under consideration. We study these abnormalities in dynamics through feature importance estimation in the resting-state functional MRI data using different interpretability methods. We validate this feature estimation by introducing the Retain and Retrain (RAR) process and demonstrate its utility on an Alzheimer's disease dataset. Furthermore, we show that the proposed model is adaptable to small sample case by offering a self-supervised pretraining scheme of the same model. With this scheme, we demonstrate that the model can leverage a large unrelated but publicly available dataset to learn improved representation to maintain adequate predictive capacity and extract useful disorder-specific information.

1 INTRODUCTION

Mental disorders are generally revealed in behavior and attributed to malfunctioning brain dynamics (Goldberg & Huxley, 1992; Calhoun et al., 2014). Understanding the signal dynamics is intuitively useful for understanding the disorder under consideration. Unfortunately, working with multivariate dynamic signals is challenging due to their high dimensionality compared to the few samples available for the study. Handcrafted summary features (Khazaee et al., 2016) may be useful to predictive models. For example, in brain imaging datasets, people frequently use correlation matrices of functional network connectivity (FNC) among different regions of the brain as proxy features (Allen et al., 2014). Unfortunately, these handcrafted features—sFNC (static FNC)—are impractical for understanding the dynamics. Yet, if traditional machine learning models directly deal with large multivariate signals, it leads to a drastic drop in performance.

Furthermore, the discovery of disorder-specific dynamics is required to forge better treatment strategies for patients. People working to explore the brain functionality related to mental disorders must know what, when and how affects spatio-temporal signals making them different from healthy people. Overall, the pragmatic necessity of discovering the essence of the underlying mechanism and the apparent impossibility of empirically dealing with extreme high dimensionality put the original learning problem in a state of quandary. Consequently, in many applications, these undesirable conditions necessitate deep learning methods leveraging their impressive ability to learn from the raw data. To this end, we propose a hierarchical recurrent neural network that performs acceptably well in terms of predictive metrics and model interpretability for multivariate time-series signals (independent component analysis (ICA) time courses (Calhoun et al., 2001).

Moreover, to enable the direct study of systems in small data studies, we propose a solution based on a self-supervised representation learning procedure. This self-supervised pre-training is useful in different computer vision (Bachman et al., 2019) and neuroimaging (Mahmood et al., 2020; 2019) applications. We leverage self-supervised pretraining guided by signal dynamics on publicly available healthy control subjects from the Human Connectome Project (HCP) (Van Essen et al., 2013).

We marshal sufficient evidence to show that the proposed deep learning model helps learn directly from dynamics and increases predictive capacity. Besides, we apply different model introspection techniques (Hooker et al., 2019) to identify important disorder-specific biomarkers supposedly suitable to advance our perception of the disorders. To evaluate the identified biomarkers, we propose a method, called RAR, suitable for ICA time courses obtained from rsfMRI data. With RAR, we show the efficacy of the identified biomarkers both in studies with or without pretraining. We verify our findings on the OASIS dataset designed for Alzheimer's research.

2 METHOD

Our method consists of 4 steps: First, we pre-train the network except for the top fully connected layer on a large unrelated and unlabeled dataset to learn a representation of the latent factors. Subsequently, we use the pre-trained weights to initialize the network during the downstream task. In the second step, we trained the downstream classification model to learn from the downstream data dynamics. In the third step, we estimated feature importance based on the model's predictions using different interpretability methods. In the fourth step, we evaluated the estimated features using a technique, called **Retain and Retrain** (RAR) as described in Section 3.3.

MILC

MILC, also called *whole MILC*, stands for "mutual information local to context." It is a self-supervised pretraining technique (Mahmood et al., 2020) used to maximize the mutual information of the latent space of a window (time slice) and the corresponding full sequence as a whole.

Let $\mathbf{D} = \{(\boldsymbol{u}_t^{\mathrm{i}}, \boldsymbol{v}^{\mathrm{j}}) : 1 \leq \mathrm{t} \leq T, 1 \leq \mathrm{i}, \mathrm{j} \leq N\}$ be a dataset of window-sequence embedding pairs computed from ICA time courses, where subscript t refers to the t-th window, superscripts i, j each refers to a sequence number. T is the number of windows in a sequence, and N is the total number of sequences in the dataset. $\mathbf{D}$ can be decomposed into a set of positive pairs $\mathbf{D}^+$ ($\mathrm{i} = \mathrm{j}$) and a set of negative pairs $\mathbf{D}^-$ ($\mathrm{i} \neq \mathrm{j}$) denoting a joint and a marginal distribution respectively for the window-sequence pairs in the latent space. With a critic function f, we use InfoNCE estimator Oord et al. (2018) to compute a lower bound $\mathcal{I}_f(\mathbf{D}^+)$ on the mutual information defined as:

$$\mathcal{I}(\mathbf{D}^+) \geq \mathcal{I}_f(\mathbf{D}^+) \triangleq \sum_{\mathrm{i}=1}^{N} \sum_{\mathrm{t}=1}^{T} \log \frac{\exp f((\boldsymbol{u}_{\mathrm{t}}^{\mathrm{i}}, \boldsymbol{v}^{\mathrm{i}}))}{\sum_{\mathrm{k}=1}^{N} \exp f((\boldsymbol{u}_{\mathrm{t}}^{\mathrm{i}}, \boldsymbol{v}^{\mathrm{k}}))}, \tag{1}$$

where f is a separable critic defined as $f(\boldsymbol{u}_{\mathrm{t}}, \boldsymbol{v}) = \phi(\boldsymbol{u}_{\mathrm{t}}^{\mathrm{i}})^{\intercal}(\boldsymbol{v}^{\mathrm{j}})$, where ϕ is some embedding function learnt by network parameters. Critic learns an embedding function such that it assigns higher values for positive pairs than for negative pairs, i.e., $f(\mathbf{D}^+) \gg f(\mathbf{D}^-)$. To make it precise, $\boldsymbol{u}_{\mathrm{t}}$ and $\boldsymbol{v}$ in the Equation 1 respectively refer to window embedding $\boldsymbol{z}_{\mathrm{t}}$ and global sequence embedding $\boldsymbol{c}$.

3 EXPERIMENTS

3.1 DATASET

We used ICA time courses computed from HCP dataset (823 healthy subjects) and OASIS (Open Access Series of Imaging Studies) (Rubin et al., 1998)(372 subjects) dataset for Alzheimer's disease respectively for the pretraining and downstream task. We received 100 ICA components using the same procedure as described in (Fu et al., 2019). However, we used only 53 non-noise features as determined per slice (time point) in all experiments. We divided the ICA time courses into windows of 20 time points for all experiments with 95% overlap (window shift = 1) along the time dimension.

3.2 WHOLE MILC SETUP

The unidirectional recurrent encoder with an attention mechanism takes 53×20 windows and represents each time point with its hidden 256-dimensional representation. We pass these hidden dimensional representations through an attention network, a two-layer feed-forward network with

hidden units 64, to produce a series of weights representatives of the required degrees of attention. The hidden representations are then weighted to produce window embedding $\boldsymbol{z}$. We initialized the LSTM and linear units in the network with Xavier initialization. We used Adam optimizer both for pretraining and training.

We used 700 and 123 HCP subjects, respectively, for pretraining and its evaluation. In MILC based pretraining, we pass encoder embeddings $\boldsymbol{z}$ to another unidirectional recurrent network with an attention mechanism. The hidden dimension for this higher (according to hierarchy) recurrent network is 200. The used attention mechanism uses 400 input neurons (because of anchoring with the final hidden state at each time point), 128 hidden units to produce a set of weights. These weights are used as coefficients in the linear combination of hidden representations to generate a global embedding $\boldsymbol{c}$ of dimension 200. Based on $\boldsymbol{c}$ and $\boldsymbol{z}$, we pre-train the neural network using the mechanism as described in Section 2. Window and sequence embeddings match with 89% accuracy during pretraining evaluation.

3.3 WHOLE MILC EVALUATION

We evaluated the effectiveness of pretraining using two downstream models—NPT and UFPT. UFPT stands for the **"Unfrozen Pre-trained"** model, for which we further train the model on top of the pre-trained weights. NPT stands for the **"Not Pre-trained"** model, for which we initialize the model with random weights and train the model with the downstream data. We progressively increase the downstream data to show the effects of pretraining. Furthermore, we compare their relative performance in terms of the available biomarkers through model interpretability.

For downstream classification of subjects into patients (AZ) and controls (HC) from the OASIS dataset, we feed the ICA time courses into the recurrent encoder and obtain $\boldsymbol{z}$ and $\boldsymbol{c}$ using the same procedure as in pretraining. Finally, we use a feed-forward network with 200 hidden units on top to perform binary classification. We gradually increase the number of supervised training subjects to observe the pretraining effect on downstream data size. For each experiment, we performed ten trials to ensure the randomization of training samples. We used 64 hold-out subjects to evaluate the model's predictive performance. The outperformance of UFPT over NPT is evident as shown in Figure 1.

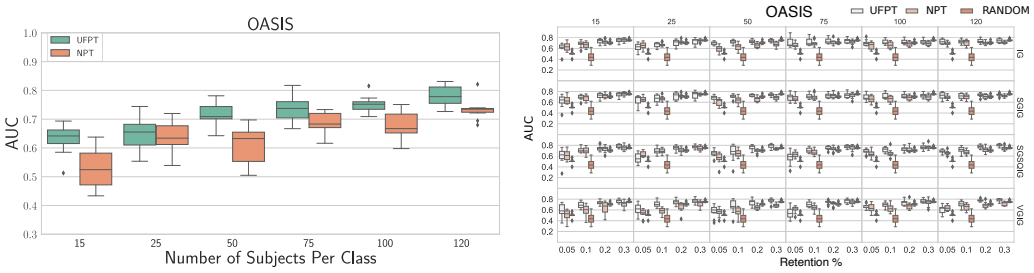

Figure 1: **Left:** It shows the main results from the whole MILC (**Model 1**). The *whole* MILC pretraining substantially improves the latent representations as reflected in the improved AUC, i.e., UFPT outperforms NPT for an equal number of subjects. However, as we hypothesized, if we gradually increase the number of subjects during training, the difference between UFPT and NPT diminishes. **Right:** RAR evaluation on SVM (**Model 2**) of different saliency methods for the OASIS dataset. Feature estimates using *integrated gradients* and its ensembles (*"smoothgrad," "smoothgrad-squared," "vargrad"*) indicate important Spatio-temporal features potentially critical to the true cause of the underlying disorders. Also, it is observable that at 5% retention of salient features, the importance of features is substantial. However, as the percentage of retention increases, random importance catches up with these estimated feature importance. We attribute this phenomenon to the redundancy in the ICA time courses data. Furthermore, the results reflect that, in the case of small data, self-supervised pretraining (UFPT) assists in estimating feature attributions more accurately than its no-pretraining (NPT) counterpart.

POSTHOC EXPLANATION AND RAR EVALUATION

After we train the *whole* MILC model and finish evaluation on the hold-out test set, we used integrated gradients (**IG**) (Sundararajan et al., 2017), and some ensembles—smoothgrad (**SG**) (Seo et al., 2018; Smilkov et al., 2017), smoothgrad-squared (**SG-SQ**), vargrad (**VG**) (Adebayo et al., 2018)—of integrated gradients to compute posthoc explanation (saliency maps) of every sample in the dataset. To evaluate the posthoc explanation, we initially investigated an existing ROAR method as used in (Hooker et al., 2019) for feature evaluation. However, this method was not suitable for ICA time courses because removing a certain percentage of entries does not necessarily cause an interpretable drop in the resulting accuracy. We attribute this phenomenon to the redundancy in the ICA time courses. Hence, we ended up with the RAR method. In RAR, as opposed to ROAR, we retain a small percentage of the features taken in descending order of importance. We compute sFNC of these modified samples, retrain and evaluate a nonlinear SVM model (**Model 2**). We use the same train-test split for the RAR evaluation as used in the whole MILC (**Model 1**). Finally, we compare the performance obtained with estimated features against the random baseline.

Let us define $\mathbf{X}$ as the original dataset (*ICA time courses of subjects*) and $\mathbf{X}^M \mid g^{\mathbf{R}}$ be the modified dataset based on random importance estimates and $\mathbf{X}^M \mid g_i$ be the modified dataset by some saliency method g_i. To guarantee that obtained saliency estimates are meaningful and have more disorder-specific information, we show that $\xi(\mathbf{X}^M \mid g_i) > \xi(\mathbf{X}^M \mid g^{\mathbf{R}}$, where ξ is the performance evaluation function, e.g. area under curve and/or accuracy.

As observed in Figure 1, the dynamics learned by the model are spatiotemporally meaningful as the small percentage of the critical features outperforms a similar amount of randomly chosen features when we evaluate them on Model 2 (**SVM**). Besides, as reflected in AUC, the biomarkers identified with UFPT models seem empirically of superior quality than its NPT counterpart. This encouraging result generalizes, in most cases, across the datasets, even when we use very few subjects (15) for training.

4 CONCLUSION AND FUTURE WORK

In this work, we propose a deep learning model that leverages its ability to learn directly from signal dynamics rather than using pre-engineered features. We empirically show that the proposed model can perform reasonably well in predictive metrics on the OASIS dataset designed for Alzheimer's study. We introspect the model for the critical but undiscovered disorder-specific information in the dynamics. To this end, we used several saliency-based interpretability methods to estimate feature attributions according to the order of importance the model assigns to them for its predictive decisions. Furthermore, we propose a posthoc explanation evaluation method called RAR, which can effectively evaluate the estimated features' significance. With RAR, we demonstrate that important features as determined by the learned model outweigh randomly selected features and thus capture significant disorder-relevant parts of the dynamics. Besides, we provide a self-supervised pretraining scheme to enable the direct investigation of system dynamics in cases where the dataset's size under consideration is insufficient for the study. Toward this goal, we show that leveraging self-supervised pretraining on a publicly available unlabeled and unrelated dataset (HCP) enables small data studies. We pragmatically show that pretraining noticeably uplifts downstream performance and extracts robust features for its predictive decisions. In the future, we expect to go beyond the feature evaluation and interpret the underlying disorder-specific biomarkers from those estimated features essential to the model for its predictions.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31:9505–9515, 2018.

Elena A Allen, Eswar Damaraju, Sergey M Plis, Erik B Erhardt, Tom Eichele, and Vince D Calhoun. Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, 24(3):663–676, 2014.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

Vince D Calhoun, Tulay Adali, Godfrey D Pearlson, and JJ Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

Vince D Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The chronnectome: time-varying connectivity networks as the next frontier in fmri data discovery. *Neuron*, 84(2):262–274, 2014.

Zening Fu, Arvind Caprihan, Jiayu Chen, Yuhui Du, John C Adair, Jing Sui, Gary A Rosenberg, and Vince D Calhoun. Altered static and dynamic functional network connectivity in alzheimer's disease and subcortical ischemic vascular disease: shared and specific brain connectivity abnormalities. *Human Brain Mapping*, 2019.

David P Goldberg and Peter Huxley. *Common mental disorders: a bio-social model.* Tavistock/Routledge, 1992.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 9737–9748, 2019.

Ali Khazaee, Ata Ebrahimzadeh, and Abbas Babajani-Feremi. Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer's disease. *Brain Imaging and Behavior*, 10(3):799–817, Sep 2016. ISSN 1931-7565. doi: 10.1007/s11682-015-9448-7. URL https://doi.org/10.1007/s11682-015-9448-7.

Usman Mahmood, Md Mahfuzur Rahman, Alex Fedorov, Zening Fu, and Sergey Plis. Transfer learning of fmri dynamics. *arXiv preprint arXiv:1911.06813*, 2019.

Usman Mahmood, Md Mahfuzur Rahman, Alex Fedorov, Noah Lewis, Zening Fu, V. D. Calhoun, and Sergey Plis. Whole milc: generalizing learned dynamics across tasks, datasets, and populations. In *Proceedings of the 23rd international conference on medical image computing and computer assisted intervention (MICCAI)*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Eugene H Rubin, Martha Storandt, J Philip Miller, Dorothy A Kinscherf, Elizabeth A Grant, John C Morris, and Leonard Berg. A prospective study of cognitive function and onset of dementia in cognitively healthy elders. *Archives of neurology*, 55(3):395–401, 1998.

Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding methods of saliency map as series of higher order partial derivative. *arXiv preprint arXiv:1806.03000*, 2018.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.