BEBOP: Bidirectional dEep Brain cOnnectivity maPping

Riccardo Asnaghi
Politecnico di Milano
Milano, Italy
riccardo1.asnaghi@mail.polimi.it

Letizia Clementi
Politecnico di Milano
CHDS, Center for Health Data Science
Human Technopole, Milan, Italy
Milano, Italy
letizia.clementi@polimi.it

Marco Domenico Santambrogio
Politecnico di Milano
Milano, Italy
marco.santambrogio@polimi.it

***Abstract*—Functional connectivity mapping provides information about correlated brain areas, useful for many applications such as on mental disorders. Thereby this work aims to improve this mapping by using deep metric learning considering the directionality of information flow and time-domain features. To deal with the computational cost of a complete pairwise combination network, we trained a network able to recognize similar signals and, after training, feed it with all combinations of signals from each brain area. The labels of similarity or dissimilarity are determined by agglomerative clustering using the Jensen-Shannon Distance (JSD) as a metric. To validate our approach we employed a resting-state eye-open functional Magnetic Resonance Imaging (fMRI) datasets from ADHD and healthy subjects. After obtaining the maps from each subject, and noticing the difference, we perform a feature importance selection using logistic regression. The ten most promising areas were extracted, such as the frontal cortex and the limbic system. These results are in complete agreement with previous literature. It is well known that the frontal cortex and the limbic system are mainly involved in attention and impulsivity.**

***Index Terms*—Deep Metric Learning, Deep Learning, ADHD, functional MRI, Functional Data Analysis**

I. Introduction

Functional connectivity mapping is the process of individuating brain areas that share a statistical relationship in time between the measures of activity recorded from them, hence that areas are presumed to be correlated [1]. It is assumed that functionally connected areas participate in the same network and pathway of information flow [1]. This means that improving the mapping of functional connectivity could lead to new insights into the role of Central Nervous System (CNS), specifically its involvement in mental illness. One of the approaches most commonly used in the clinic is the seed-based method. In this technique, the correlation between one brain area and the other voxels is calculated using Pearson correlation or Mutual Information (MI). This method is fast and simple, however, it does not provide satisfactory results for complex problems, such as highly noisy fMRI signals at rest, and it is heavily dependent on the chosen Region Of Interest (ROI).

In this project, we aim to improve functional connectivity mapping methods by using Deep Learning (DL) and Functional Data Analysis (FDA), as preprocessing tool, to evaluate the directionality of information flow and give more relevance to features in the time domain. Specifically, we use a metric learning network, to create a representation of the inputs and then evaluate the similarity using a distance metric [2]. The use of DL allows us to process the noise of resting-state fMRIs through convolutional layers that embed the signal in a more informative way [3], [4]. Instead, Recurrent Neural Network (RNN) takes into account time flow thanks to their cyclical connection that allows them to update the state sample by sample [5]. To better estimate the average response from each brain area, we employ FDA, a branch of statistics that analyzes data in a continuum space by interpolating or smoothing discrete data and converting them into analytic functions. This approach is particularly useful because it allows us to analyze morphological trajectories and time delays as well as shifts [6], [7]. Specifically, FDA allows us to assess a morphological distribution of curves, and remove outliers based on the whole shape of the signal [8]. These will be the inputs used by our network to compute the directional flow of information. This approach is validated through the analysis of fMRI data from healthy and Attention Deficit Hyperactivity Disorder (ADHD) subjects at rest since altered brain activity and connectivity are well-studied [9]–[11].

II. Material & Methods

A. Dataset

In the present work, we exploit the data from the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study [12], a large dataset of partially preprocessed 3T fMRI with a TR of 2s from 290 subjects between 25 and 50 years of age. It includes four different subgroups of subjects and seven experimental protocols. The population includes healthy subjects (130 sbj.) and three pathological groups: ADHD (43 sbj.), bipolar disorder (49 sbj.), and schizophrenia (50 sbj.). The performed protocols consist of a resting-state, a balloon analog risk task, a paired-associate memory task, a spatial working memory task, a stop-signal task, a task-switching task, and a breath-holding task. In particular, we consider 5-minute eye-open resting-state fMRI scans of ADHD and healthy subjects.

B. Registration and Data Preprocessing

The analysis begins with multimodal affine registration on an asymmetric Hammer atlas that parcellates the brain into 83 regions [13], [14]. This procedure allows the images to be aligned on the atlas and allows us to locate signals to specific areas. Since the images and the template comes from different sources and there is no direct correlation between the gray levels, MI is used as a nonlinear metric. The affine registration allows the transformation of the scans with twelve degrees of freedom by translation, rotation, shear, and scaling. Then, each signal is filtered between 0.01 and 0.16 Hz to partially remove noise using a 4th order Butterworth filter forward and backward. A functional trimmed mean based on the depth measure is applied to the signals from each area. This method is an extension of the trimmed mean for functional data based on the concept of depth in FDA, a measure of outlyingness, and it consists of calculating the mean function excluding the least deep curves [15]. Depth measures are indices that assign a value to each function indicating how far it is from the curve distribution center, and the maximum depth corresponds to the median of the distribution [15]. This approach allows us to consider the distribution of the BOLD signals of a brain region and produce a reliable summary signal from each subject that is less biased by noise because outliers are removed.

C. Hierarchical Clustering

A naive idea could be to train a single classification network that can combine the inputs from each brain in pairs and then classify the subject according to its group based on the bidirectional neighboring matrix. However, from a computational point of view, this approach is impractical, since the problem scales quadratically with inputs (83 inputs regions imply 6889 pairwise branches). To reduce the computational complexity, we train a network capable of classifying similarities between a pair of signals and extracting a bidirectional score. To train this type of network, we use similarity labels. We compute them through hierarchical agglomerative clustering with the Jensen-Shannon Distance (JSD) to extract the 1000 most correlated signals without considering the location in the brain. This metric is defined as: $JSD(p,q) = \sqrt{\frac{D(p\|m)+D(q\|m)}{2}}$ where $D(...)$ is the Kullback-Leibler divergence with respect to the pointwise mean of the distribution p and q. It allows us to compute a nonlinear correlation between areas since the distance is calculated on distributions and not on the signal [16]. Before the clustering process, a *robust scaler* is applied to make the samples comparable and improve the clustering performance. Thus, each point is scaled by subtracting the median of the point and dividing by the interquartile range.

D. The Network

The network aims at extracting the directional flow of information between two different inputs. Thus, it must be able to distinguish the information from the first input to the second and vice versa. Our system consists of three blocks: encoding block, similarity block, and classification block. The network's parameters are optimized through a Bayesian optimization tuner [17]. It is important to highlight that the first two blocks share weights, as in a metric learning approach, to create a joint representation of the two inputs. These joint representations are needed to compute the distance between them and estimate the similarity. Figure II-D displays a graphical representation of the network. A detailed explanation of each block is provided below.

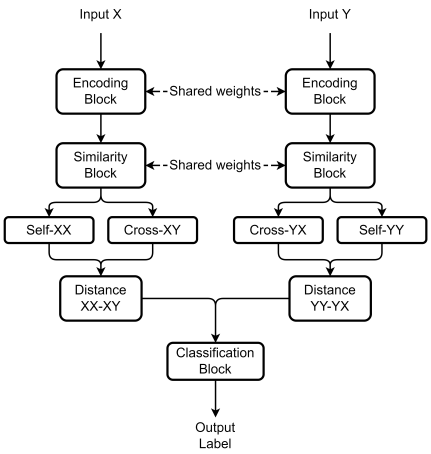

Fig. 1. Complete overview of our network. After the encoding and the similarity block, where weights are shared as in a metric learning approach, a distance between representations is computed. The distance is then used to classify the similarity between the input signals or not.

1) Encoding block: Encoding the inputs allows extracting features to be inserted into the following block. After a simple feed-forward layer to increase the initial depth, the input first passes through 2 encoding blocks. Each block encodes time through a Gated Recurrent Unit (GRU) [18] bidirectional layer [19], then features are extracted with a ResNet-like [20] structure built by a causal convolution layer [21]. Causal convolution is a type of convolution that preserves the order of sample, which can be seen as a convolution previously multiplied by a triangular filter [21]. Since this layer cannot violate the time-flow direction, features are expected to be more sensitive to time. Finally, the result undergoes layer normalization [22], and max-pooling is applied to increase the field of view of the network and decrease the width. Figure 2 shows in detail the architecture and the parameters used.

2) Similarity block: This block represents the core of our network. It aims at computing the distance between representations, as in metric learning. By using the Multi-Head attention layer [23], we can construct four different representations: one for each combination of the two inputs. Two of these are called self-representation because the query and key entries in the Multi-Head attention layer are the same, and the remaining two are called cross-representations because the query and key are different. Using Euclidean distance, the network calculates the distance between similar self- and cross-representations. Since only the query value changes between representations (the weights are shared between similar layers), we can assume

that the information flowing in a given direction from one signal to the other is proportional to the distance between the generated self and cross representations. In this way, the calculated attention becomes an index of directional similarity. An overview of this block, with all the parameters, is reported in Figure 2.

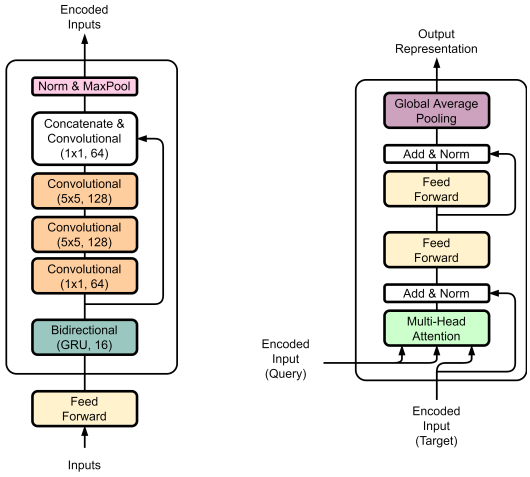

Fig. 2. Overview of encoding and similarity block of the net. Feedforward layer encodes the inputs (shape: [152,1] each) with 64 neurons. The results pass through 2 blocks of encoding, the parameters are reported in the figure. Before the similarity block, a bidirectional layer with 32 neurons fix the channel dimension, also for the following layers. Multi-Head attention layers with a key dimension of 80 and 24 heads compute the similitude between key and query. All activation functions are *ReLU*, except for bidirectional layers that use *tanh* and sigmoid.

3) Classification block: The distances between representations are the inputs of the classification layer once they are concatenated. Then a feed-forward layer with two neurons with a linear activation mixes the results and finally, a single neuron with a sigmoidal activation classifies the signal by similarity or not. This last block is removed after the training, once its purpose is fulfilled. Namely, our goal is to create an adjacency matrix that is filled by the distance between the representations that represent the flow of information.

E. Results evaluation

After training, an adjacency matrix is generated combining the 83 inputs in pairs, one for each region. We could not evaluate by classification between healthy and ADHD patients could not be performed due to the curse of dimensionality [24]. Indeed, the number of features is much higher than the number of samples. Therefore, we perform feature selection using logistic regression. We select the 10 most promising features and evaluate our results by comparing them with previous literature.

III. Results

The network is trained as a classification task, with binary cross-entropy as loss and accuracy as metric, using Adam [25] with an initial learning rate of 10^{-5}. The training also involves an early stopping to prevent overfitting and a learning rate reduction when a plateau of the metric is reached. As result, the 98% accuracy is reached in 167 epochs.

Once the network is trained, the classification block is removed and the two distances are extracted from each pair of signals. In this way, we obtain a connectivity map for each patient from both healthy and ADHD subjects. To be comparable, we apply a within-subjects *min-max* scaling; hence, each map is scaled according to the minimum and maximum of each subject. Then, we average together the maps following the group they belong. Figure 3 summarizes the average results.

As previously explained, a classification task on the maps cannot be performed due to the curse of dimensionality [24]. For this reason, the ten most correlated features, five positives, and five negatives are extracted. Table I summarize these areas. Positively correlated features are mostly related to ADHD subjects while negative to CTR subjects since in our analysis ADHD is coded as class (*1*) while healthy to class (*-1*).

TABLE I
FEATURE IMPORTANCE SELECTION BASED ON LOGISTIC REGRESSION RESULTS.

	MOST CORRELATED AREAS		
	From:		**To:**
Positive	Anterior temporal lobe, lateral part, S	→	Parahippocampal and ambient gyri, D
	Anterior temporal lobe, lateral part, S	→	Middle and inferior temporal gyrus, S
	Middle and inferior temporal gyrus, S	→	Hippocampus, S
	Hippocampus, S	→	Parahippocampal and ambient gyri, D
	Hippocampus, S	→	Middle and inferior temporal gyrus, S
	From:		**To:**
Negative	Pre-subgenual frontal cortex, S	→	Lateral orbital gyrus, D
	Pre-subgenual frontal cortex, S	→	Lateral orbital gyrus, S
	Putamen, D	→	Cerebellum, S
	Precentral gyrus, D	→	Lateral ventricle (excluding temporal horn), S
	Cuneus, D	→	Pre-subgenual frontal cortex, D

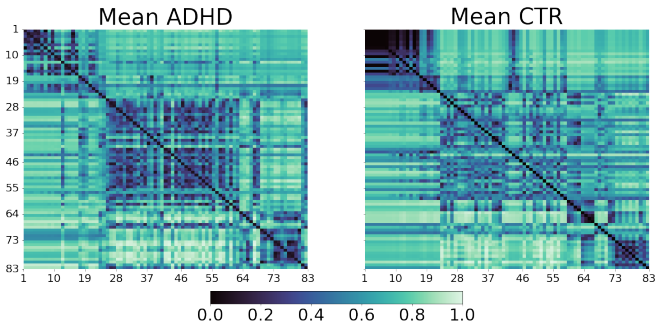

Fig. 3. Average correlation map for both ADHD and healthy subjects. On both axes, the number of each Hammer's area is reported. Colorbar displays the similarity score which ranges from 0 to 1.

IV. Discussion

In the present work, we use DL for mapping the connectivity between brain areas through the recognition of similar signals, one from each area due to the distribution of curves extracted with FDA methods. As shown in Figure 3, the average outcomes of the two groups after normalization are quite different. A different activation pattern can be seen in the top right and the bottom left area of the map. These findings are confirmed by the results of the feature selection applied through logistic regression. The areas involved, as reported in Table I, belong to four main sites: the frontal cortex, the limbic system,

the cerebellum, and the basal ganglia. According to previous literature, these areas are among the most involved in the ADHD disturb [26]. The frontal cortex is related to attention mechanisms and organization [26], [27] suggests that this area is impaired in subjects affected by ADHD. Conversely, the limbic system, cerebellum, and basal ganglia are associated with behavior, emotions, and impulsivity. An impairment of these areas could explain the symptoms of ADHD subject [26], [28]. Subjects affected by ADHD are characterized by, apart from the concentrating issues, restlessness, and impulsivity.

The main limitation of our approach is the computational cost concerning the gold standard. On contrary, our method is highly non-linear and includes a preprocessing part. No noise reduction is manually performed on data, except for the initial filtering. The network is freely left to learn a latent representation useful for similarity recognition. Another main advantage is the complete model-agnosticism since no assumptions about the model itself are made.

V. Conclusion

In this work, we present a method for investigating functional connectivity based on DL supported by FDA. The developed network, inspired by metric learning, aims at evaluating the similarity and directionality between regions using a distance metric. In addition, the network is built to give more relevance to time features through the use of bidirectional layers and causal convolution. Results on ADHD and healthy subjects suggest an initial step in that direction. The most involved areas, extracted through logistic regression feature selection, are the ones correlated with the pathology considered. Further investigation of this approach could lead to a new way of looking at functional connectivity; not only for the evaluation of the directionality of the information but also for its robustness due to the encoding with a convolutional layer.

Future works will include a more detailed investigation of the method such as an evaluation of other distance metrics, as well as an application of the method to another subgroup of the same dataset such as bipolar disorder. Another interesting point could be the implementation of graph deep learning methods to extract only the informative network from the generated adjacency matrix. Finally, due to the intrinsic model-agnosticism of the method, we could try to adapt the domain to other signals such as EEG.

References

[1] B. Biswal, F. Zerrin Yetkin, V. M. Haughton, and J. S. Hyde, "Functional connectivity in the motor cortex of resting human brain using echo-planar mri," *Magnetic Resonance in Medicine*, no. 4, pp. 537–541.

[2] L. Yang and R. Jin, "Distance Metric Learning: A Comprehensive Survey, Michigan State Universiy," *Technical reports, Michigan State Universiy*, p. 4.

[3] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep Learning is Robust to Massive Label Noise."

[4] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May. Institute of Electrical and Electronics Engineers Inc., may 2016, pp. 5695–5699.

[5] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," pp. 1235–1270, jul.

[6] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, ser. Springer Series in Statistics. New York, NY: Springer New York.

[7] P. Kokoszka and M. Reimherr, *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, sep.

[8] L. Xu and Y. Hong, *Functional and Shape Data Analysis*, 2017, vol. 49, no. 4.

[9] K. Konrad and S. B. Eickhoff, "Is the ADHD brain wired differently? A review on structural and functional connectivity in attention deficit hyperactivity disorder," pp. 904–916, jun.

[10] C. Chen, H. Yang, Y. Du, G. Zhai, H. Xiong, D. Yao, P. Xu, J. Gong, G. Yin, and F. Li, "Altered Functional Connectivity in Children with ADHD Revealed by Scalp EEG: An ERP Study," *Neural Plasticity*.

[11] J. F. Lubar, "Discourse on the development of EEG diagnostics and biofeedback for attention-deficit/hyperactivity disorders," *Biofeedback and Self-Regulation*, no. 3, pp. 201–225, sep.

[12] K. J. Gorgolewski, J. Durnez, and R. A. Poldrack, "Preprocessed Consortium for Neuropsychiatric Phenomics dataset," *F1000Research*, p. 1262, sep.

[13] A. Hammers, R. Allom, M. J. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks, and J. S. Duncan, "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe," *Human Brain Mapping*, no. 4, pp. 224–247, aug.

[14] I. S. Gousias, D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers, "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest," *NeuroImage*, no. 2, pp. 672–684, apr.

[15] S. L. Pez-Pintado and J. Romo, "On the concept of depth for functional data," *Journal of the American Statistical Association*, no. 486, pp. 718–734, jun.

[16] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo, "The Jensen-Shannon divergence," *Journal of the Franklin Institute*, vol. 334, no. 2, pp. 307–318, mar 1997.

[17] C. G. E. Boender and J. Mockus, *Bayesian Approach to Global Optimization: Theory and Applications.* Springer Netherlands, 1991, vol. 56, no. 194.

[18] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST 2014 - 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics (ACL), sep, pp. 103–111.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem. IEEE Computer Society, dec 2016, pp. 770–778.

[21] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," sep.

[22] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," jul.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, jun, pp. 5999–6009.

[24] R. Bellman, "Dynamic Programming: A Rand Corporation Research Study," p. 363, 1957.

[25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec.

[26] S. Durston, "A review of the biological bases of ADHD: What have we learned from imaging studies?" *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 9, no. 3, pp. 184–195, 2003.

[27] D. Amen and B. Carmichael, "High-Resolution Brain SPECT Imaging in ADHD," *Annals of Clinical Psychiatry*, vol. 9, no. 2, pp. 81–86, jun 1997.

[28] M. M. Bruchhage, M. P. Bucci, and E. B. Becker, "Cerebellar involvement in autism and ADHD," in *Handbook of Clinical Neurology*. Elsevier, jan 2018, vol. 155, pp. 61–72.