Explaining ANN-modeled fMRI Data with Path-Weights and Layer-Wise Relevance Propagation

José Diogo Marques dos Santos[1,2], José Paulo Marques dos Santos[3,4,5*]

[1] Faculty of Engineering, University of Porto, R. Dr. Roberto Frias, 4200-465 Porto, Portugal
[2] Abel Salazar Biomedical Sciences Institute, University of Porto, R. Jorge de Viterbo Ferreira, 4050-313 Porto, Portugal
[3] University of Maia, Av. Carlos de Oliveira Campos, 4475-690 Maia, Portugal
[4] LIACC - Artificial Intelligence and Computer Science Laboratory, University of Porto, R. Dr. Roberto Frias, 4200-465 Porto, Portugal
[5] University of Porto, Faculty of Medicine, Unit of Experimental Biology, Alameda Prof. Hernâni Monteiro, 4200-319 Porto, Portugal

Abstract

It may be possible to extract knowledge from functional magnetic resonance (fMRI) data with artificial neural networks (ANNs) and explainable artificial intelligence (xAI). However, modeling fMRI data with ANNs has its hurdles. One is the unbalance between inputs (one typical volume encompasses hundreds of thousands of voxels) and training epochs (usually hundreds), turning the training stage intractable. In addition, fMRI data is noisy and highly correlated, both spatially and temporally. Such characteristics tend to hamper current deep learning techniques and, therefore, limit fMRI data modeling and their explanation.

The research here reported relies on a process encompassing data splitting by training and testing, dimensionality reduction, feature extraction, ANN structuring, and its training and testing. After the procedure, two explaining methods are put side by side, path-weights and layer-wise relevance propagation (LRP).

The two methods achieve similar results, i.e., identify the same inputs responsible for the ANN's correct predictions. Therefore, they support each other. An additional validation comes from neuroscientific established knowledge, which sanctions the results of the two explaining methods. A publicly accessible database, Human Connectome Project (HCP) – Young Adults, precisely the motor paradigm, is used to apply the procedure.

In conclusion, the combined use of XAI techniques with ANNs modeling permits the extraction of knowledge from fMRI data, at least concerning motor tasks. The complete procedure is an improvement over traditional data analysis methods, which are correlational. The following steps will extend the procedure to cognitive tasks.

Keywords

Artificial neural networks (ANN), Explainable artificial intelligence (XAI), Layer-wise relevance propagation (LRP), Functional magnetic resonance imaging (fMRI)

1. Introduction

Although functional magnetic resonance imaging (fMRI) data analysis with artificial neural networks (ANNs) is more than one decade old [1-6], it has received renewed recent interest [7-9]. Inherently noisy and highly correlated data, unbalance between the number of inputs and training epochs, and difficulty in finding pertinent features are some of the common hurdles. An additional complication in ANNs is understanding how they make predictions [10-12]. Models that deliver high prediction accuracies are helpful. However, if they are transparent, allowing one

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

* Corresponding author.
up201908014@edu.fe.up.pt (J. D. Marques dos Santos)
jpsantos@umaia.pt (J. P. Marques dos Santos)
0000-0003-4101-2748 (J. D. Marques dos Santos)
0000-0002-5567-944X (J. P. Marques dos Santos)

to understand which inputs contribute more to the correct hits (explain) and understand how the progress of calculation leads to the prediction (interpret), they would be even more helpful. Therefore, explainable and interpretable artificial intelligence (XAI) in ANNs is needed in neuroscience.

Addressing the explainability of ANN-built models of fMRI data has been recently tackled [13-15]. One computational model for such purpose is layer-wise relevance propagation (LRP) [16, 17], which was already applied in ANNs [7, 18]. The purpose of the present study is to put side-by-side LRP and the path-weights concept suggested in [13, 14], answering the question: is the path-weights-based analysis in accordance with the state-of-the-art method of LRP-based explainability regarding input importance for the network's prediction?

2. Method

The fMRI data processing stages are represented in Figure 1: firstly, the raw data processing; next, the two datasets, train and test, are used to build the model (ANN); and finally, the model is explained with path-weights and LRP.

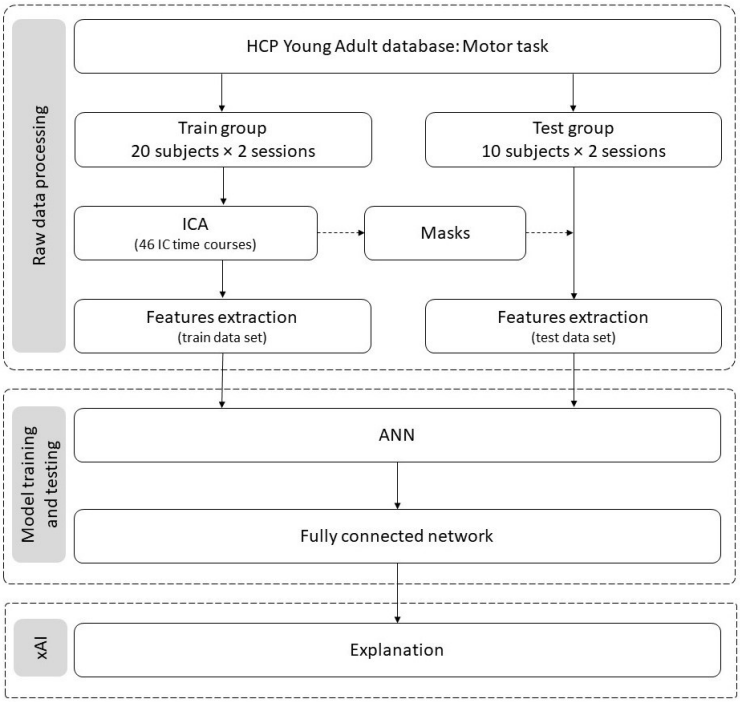

Figure 1: Flowchart of the global procedure.

2.1. Raw Data Processing and Model Training and Testing

The first two stages were already described and discussed elsewhere [13, 14]. The raw data is obtained from the publicly accessible website of the Young Adults database of the Human Connectome Project (HCP), motor paradigm in the 100 Unrelated Subjects subset [19-22].

The ANN has one hidden layer composed of 10 hidden nodes. Inputs are 46, and the outputs are five, each corresponding to a task (LF, LH, RF, RH, and T).

2.2. Explaining with Path-Weights and Layer-Wise Relevance Propagation

The $path\text{-}weight_{ijk}$ is defined as the module of the product of all connection weights in a path defined from the input I_i to the output O_k, passing by the hidden node H_j [13]:

$$path\text{-}weight_{ijk} = \left| w_{I_i H_j} \times w_{H_j O_k} \right| \quad (1)$$

where $w_{I_i H_j}$ is the weight between the input node I_i and the hidden node H_j, and $w_{H_j O_k}$ is the weight between the hidden node H_j and the output node O_k.

LRP is a state-of-the-art method for explaining ANN's predictions [16]. It is calculated according to the basic rule (LRP-0) [17]:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k \quad (2)$$

where R_j is the relevance of node j, a_j is the activation of node j, w_{jk} is the weight of the connection between nodes j and k, and R_k is the relevance of node k.

LRP computation is implemented using R's library innsight [23], version 0.2.0.

2.3. Grand-Weight (GW) and Grand-Relevance (GR) Computation

There is a need for a metric that allows for the direct comparison of individual inputs between themselves for a given stimulus. In this way, the metric Grand-weight (GR) is the sum of the absolute values of the path-weights from each input to a specific output, according to the formula:

$$GW_{ik} = \sum_j \left| path\text{-}weight_{ijk} \right| \quad (3)$$

where GW_{ik} is the Grand-weight from input i to output k, and $path\text{-}weight_{ijk}$ is the path-weight that goes from input i to output k through hidden node j.

For the LRP-based analysis, to keep consistency with the path-weights-based analysis, the absolute values of the relevancy score for each input for a given output are summed, obtaining the metric Grand-relevance (GR), which allows the direct comparison between inputs regarding their relevance. GR formula is:

$$GR_{ik} = \sum_l |R_{ikl}| \quad (4)$$

where GR_{ik} is the Grand-relevance from input i to output k, and R_{ikl} is the relevance score for input i for output k for computational epoch l.

3. Results

Networks' partial and global accuracies and precisions are represented in Table 1.

Table 1

Confusion matrix of the ANN predictions, including the partial and global accuracies and precisions (LF: left foot; LH: left hand; RF: right foot; RH: right hand; T: tongue).

Stimulus		Prediction					Total
		LF	LH	RF	RH	T	
Input	LF	27	1	6	5	1	40
	LH	3	36	0	1	0	40
	RF	8	0	31	1	0	40
	RH	0	2	1	37	0	40
	T	4	0	1	0	35	40
Total		42	39	39	44	36	200
Accuracy (%)		67.5	90.0	77.5	92.5	87.5	83.0
Precision (%)		64.3	92.3	79.5	84.1	97.2	

Figure 2 and Figure 3 depict the path-weights and the LRP calculated for the ANN. In both cases, the top five most relevant inputs for all the stimuli are IC 5, IC 7, IC 11, IC 12, and IC 14, although the magnitudes are not constant across all stimuli.

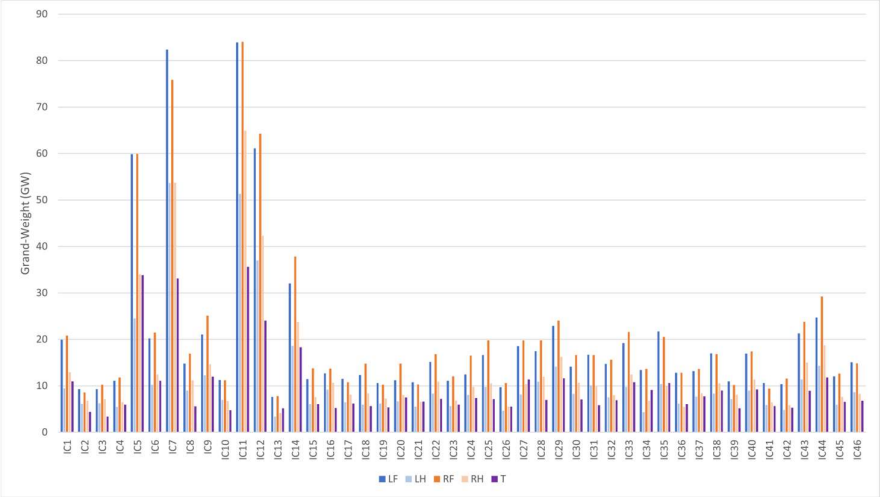

Figure 2: Grand-weight (GW) values per output for the 46 inputs.

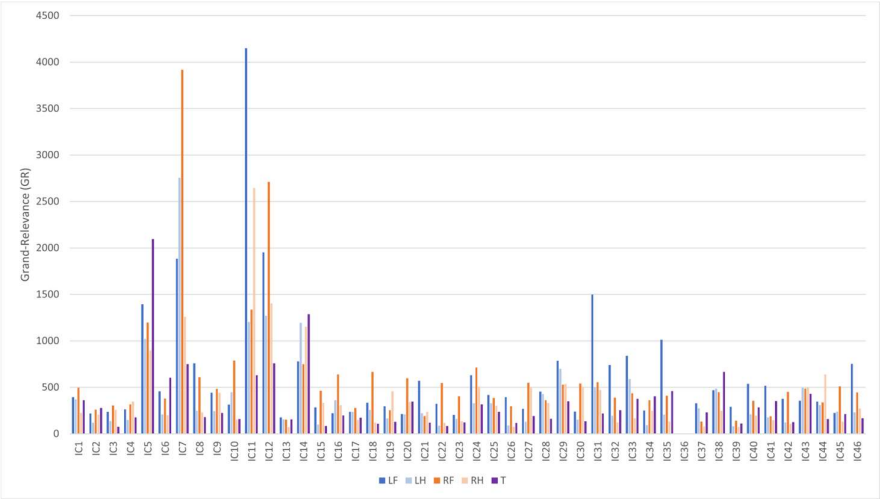

Figure 3: Grand-relevance (GR) values per output for the 46 inputs.

4. Discussion

Overall, both methods yield the same inputs as the most important. So, the path-weights-based analysis is congruent with the state-of-the-art method, LRP. Hence, it is possible to conclude that the path-weight-based procedure explains the ANN model in coherence with the layer-wise relevance propagation-based method.

Acknowledgments

This work was partially financially supported by Base Funding - UIDB/00027/2020 of the Artificial Intelligence and Computer Science Laboratory – LIACC - funded by national funds through the FCT/MCTES (PIDDAC).

References

[1] S. J. Hanson, T. Matsuka, J. V. Haxby, Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area?, NeuroImage 23 1 (2004) 156-166. doi: 10.1016/j.neuroimage.2004.05.020.

[2] M. Misaki, S. Miyauchi, Application of artificial neural network to fMRI regression analysis, NeuroImage 29 2 (2006) 396-408. doi: 10.1016/j.neuroimage.2005.08.002.

[3] D. Sona, S. Veeramachaneni, E. Olivetti, P. Avesani, Inferring cognition from fMRI brain images, in: J. Marques de Sá, L. Alexandre, W. Duch, D. Mandic (Eds.), Artificial Neural Networks – ICANN 2007, volume 4669 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2007, pp. 869-878. doi: 10.1007/978-3-540-74695-9_89.

[4] R. Espírito-Santo, J. R. Sato, M. d. G. M. Martin, Discriminating brain activated area and predicting the stimuli performed using artificial neural network, Exacta 5 2 (2007) 311-320. doi: 10.5585/exacta.v5i2.1180.

[5] J. P. Santos, L. Moutinho, Tackling the cognitive processes that underlie brands' assessments using artificial neural networks and whole brain fMRI acquisitions, in *Proceedings of the 2011 IEEE International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, IEEE Computer Society, Seoul, Republic of Korea, 2011, pp. 9-12. doi: 10.1109/PRNI.2011.22.

[6] C. D. Hacker, T. O. Laumann, N. P. Szrama, A. Baldassarre, A. Z. Snyder, E. C. Leuthardt, M. Corbetta, Resting state network estimation in individual subjects, NeuroImage 82 (2013) 616-633. doi: 10.1016/j.neuroimage.2013.05.108.

[7] A. W. Thomas, H. R. Heekeren, K.-R. Müller, W. Samek, Analyzing neuroimaging data through recurrent deep learning models, Frontiers in Neuroscience 13 (2019). doi: 10.3389/fnins.2019.01321.

[8] A. W. Thomas, C. Ré, R. A. Poldrack, Interpreting mental state decoding with deep learning models, Trends in Cognitive Sciences 26 11 (2022) 972-986. doi: 10.1016/j.tics.2022.07.003.

[9] M. Liu, R. C. Amey, R. A. Backer, J. P. Simon, C. E. Forbes, Behavioral studies using large-scale brain networks – Methods and validations, Frontiers in Human Neuroscience 16 (2022). doi: 10.3389/fnhum.2022.875201.

[10] W. Samek, K.-R. Müller, Towards explainable artificial intelligence, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, volume 11700 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2019, pp. 5-22. doi: 10.1007/978-3-030-28954-6_1.

[11] A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access 6 (2018) 52138-52160. doi: 10.1109/ACCESS.2018.2870052.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 5 (2018) 1-42. doi: 10.1145/3236009.

[13] J. D. Marques dos Santos, J. P. Marques dos Santos, Towards XAI: Interpretable shallow neural network used to model HCP's fMRI motor paradigm data, in: I. Rojas, O. Valenzuela, F. Rojas, L. J. Herrera, F. Ortuño (Eds.), Bioinformatics and Biomedical Engineering, volume 13347 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2022, pp. 260-274. doi: 10.1007/978-3-031-07802-6_22.

[14] J. D. Marques dos Santos, J. P. Marques dos Santos, Path weights analyses in a shallow neural network to reach Explainable Artificial Intelligence (XAI) of fMRI data, in: G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. Pardalos, G. Di Fatta, G. Giuffrida, R. Umeton (Eds.), Machine Learning, Optimization, and Data Science, volume 13811 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2023, pp. 417-431. doi: 10.1007/978-3-031-25891-6_31.

[15] A. W. Thomas, C. Ré, R. A. Poldrack, Benchmarking explanation methods for mental state decoding with deep learning models, NeuroImage 273 (2023) 120109. doi: 10.1016/j.neuroimage.2023.120109.

[16] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE 10 7 (2015) e0130140. doi: 10.1371/journal.pone.0130140.

[17] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: An overview, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, volume 11700 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2019, pp. 193-209. doi: 10.1007/978-3-030-28954-6_10.

[18] I. Sturm, S. Lapuschkin, W. Samek, K.-R. Müller, Interpretable deep neural networks for single-trial EEG classification, Journal of Neuroscience Methods 274 (2016) 141-145. doi: 10.1016/j.jneumeth.2016.10.008.

[19] D. C. Van Essen, M. F. Glasser, The Human Connectome Project: Progress and Prospects, Cerebrum: the Dana Forum on Brain Science 2016 (2016) cer-10-16. doi:

[20] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, K. Ugurbil, The WU-Minn Human Connectome Project: An overview, NeuroImage 80 (2013) 62-79. doi: 10.1016/j.neuroimage.2013.05.041.

[21] J. S. Elam, M. F. Glasser, M. P. Harms, S. N. Sotiropoulos, J. L. R. Andersson, G. C. Burgess, S. W. Curtiss, R. Oostenveld, L. J. Larson-Prior, J.-M. Schoffelen, M. R. Hodge, E. A. Cler, D. M. Marcus, D. M. Barch, E. Yacoub, S. M. Smith, K. Ugurbil, D. C. Van Essen, The Human Connectome Project: A retrospective, NeuroImage 244 (2021) 118543. doi: 10.1016/j.neuroimage.2021.118543.

[22] K. Ugurbil, D. C. Van Essen, Human Connectome Project, Young Adult database, 2017. url: https://www.humanconnectome.org/study/hcp-young-adult.

[23] N. Koenen, R. Baudeu, innsight: Get the Insights of your Neural Network (0.2.0), 2023. url: https://cran.r-project.org/package=innsight.