Understanding the Importance of Brain Subnetworks with Shapley Values During Narrative Processing

Aurora Rossi, Yanis Aeschlimann, Emanuele Natale, Samuel Deslauriers-Gauthier, Peter Ford Dominey

Understanding the Importance of Brain Subnetworks with Shapley Values During Narrative Processing

Aurora Rossi[1*], Yanis Aeschlimann[2], Emanuele Natale[1], Samuel Deslauriers-Gauthier[2†], and Peter Ford Dominey[34†]

[1] COATI, Université Côte d'Azur, INRIA, CNRS, I3S, Sophia Antipolis, France,
[2] CRONOS, Inria Centre at Université Côte d'Azur, Sophia Antipolis, France,
[3] INSERM UMR1093-CAPS, Université Bourgogne Franche-Comté, UFR des Sciences du Sport, Dijon, France,
[4] Robot Cognition Laboratory, Marey Institute Dijon, France
[*] Corresponding author: aurora.rossi@inria.fr
[†]Equal contribution

1 Introduction

Narratives are sequences of events involving one or more characters, where the meaning of each event is shaped by its relationship to others within the story. Unlike static collections of information, narratives evolve over time, requiring the brain to continuously integrate and update its understanding as the story progresses. Common examples of narratives include spoken and written stories, as well as movies, all of which engage the brain in dynamic processing. Studying narratives is crucial because they are central to human communication and cognition. They reflect and simulate real-world experiences and offer a rich context for understanding how the brain handles information that is temporally and contextually dynamic, providing insights into cognitive processes [11].

We model the brain's dynamic functional connectivity during narrative processing recorded using functional Magnetic Resonance Imaging (fMRI) as temporal brain networks. In this representation, nodes correspond to brain regions, and the edges capture the evolving correlations between these regions over time [8, 4]. This approach allows us to capture how interactions among brain regions change over time as the narrative unfolds.

In this work, we propose a machine learning model designed to classify this data in a supervised setting. The model integrates a convolutional neural network (CNN) layer with a multi-layer perceptron (MLP) to classify narratives based on modality (audio or video) and content (airport or restaurant situations). Using Shapley values, we delve into the model's decision-making process and quantify the contributions of different subnetworks in the Yeo 7-subnetwork parcellation. This approach enables us to identify the most involved subnetworks in narrative tasks. Our work is the first in applying this method to the functional aspects of the brain, offering novel insights validated by existing literature [1]. This combination of established techniques advances our understanding of how the brain processes complex, dynamic information and contributes to a broader comprehension of narrative cognition.

2 Methods

Data The fMRI data used in our analysis was taken from the dataset by Nastase et al. [3]. In this dataset, 31 participants were exposed to a total of 16 three-minute stories (4 per run over 4 runs). The stories can be categorized based on both content and modality. In terms of content, there were 8 stories related to airports and 8 related to restaurants. Regarding modality, 8 stories were presented in audio format, while the other 8 were presented as movies. Within each run, participants experienced 2 movies and 2 audio stories. The dataset is balanced across modalities and content. Preprocessing converted BOLD signals into temporal graphs. We reduced motion artifacts through linear regression on movement parameters and applied a bandpass filter ($0.01 - 0.08$ Hz) to remove respiratory and cardiac noise [10]. We used the Schaefer brain atlas to define 100 ROIs, averaging BOLD time series within gray matter regions [5]. A sliding window approach divided the data into time steps. Pearson correlations between ROI time series within each window were computed, forming adjacency matrices that represent the temporal brain networks.

Model Our model takes as input a temporal brain network, represented as a three-dimensional tensor $X \in [-1,1]^{R\times R\times T}$, where R is the number of brain regions (100) and T is the number of time steps (8). The model architecture is composed by single-layer 3D CNN, followed by a max pooling layer and a MLP for classification. The CNN filter has size (R,R,τ) and captures temporal features by moving along the temporal axis. Formally, the output of the CNN layer is defined as $Y_{k,c} = \sigma(X * W + b)_{k,c} = \sigma(\sum_{i=1}^{R}\sum_{j=1}^{R}\sum_{p=1}^{\tau} X_{i,j,k+p-1}\cdot W_{i,j,p,c} + b_{k,c})$ where $Y \in \mathbb{R}^{K\times C}$ is the output tensor, $W \in \mathbb{R}^{R\times R\times\tau\times C}$ is the learnable filter tensor, $b \in \mathbb{R}^{K\times C}$ is the bias matrix and C is the number of output channels. The operations $\cdot, +$ and σ, which represents the $\mathrm{ReLU}(x) = \max(\{0,x\})$ activation function, are applied component-wise. The output tensor is then passed through a max pooling layer so that the output vector $Z \in \mathbb{R}^C$ is defined as $Z = \max_k Y[k,c]$. Finally, the output passed through a MLP of three fully connected layers with ReLU activation functions.

Shapley values Shapley values measure each player's contribution in cooperative game theory and are now adopted in machine learning to explain model predictions [6, 2]. We use Shapley values to assess the impact of specific brain subnetworks on our model's predictions. Because of the limited number of brain subnetworks defined by the 7 Yeo parcellation method [9], we can compute the exact Shapley values. The Shapley value for a brain subnetwork i is given by:

$$\phi_i(v) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(v(S\cup\{i\}) - v(S))$$

where N is the set of brain subnetworks, v is the accuracy of our model when considering the set S of brain subnetworks. To isolate the brain subnetworks in the temporal brain network X we set the entries of the other subnetworks to zero. The Shapley value $\phi_i(v)$ is the average marginal contribution of the brain subnetwork i over all possible combinations of brain subnetworks, the higher the Shapley value, the more important the brain subnetwork is for the prediction of the model.

3 Results

Experiments were performed to determine if the temporal brain networks can be used to discriminate brain functional connectivity patterns in response to audio vs. movie narratives (*Modality classification*) and airport vs. restaurant situations (*Content classification*).

The accuracy of our model excels in modality classification with a value of $\mathbf{96.32\% \pm 1.36\%}$. Content classification accuracy is slightly lower at $\mathbf{80.9\% \pm 1.75\%}$, which reflects the greater difficulty of this task. To evaluate the importance of temporal dynamics, we permuted the time steps of the brain networks and retrained the model. The results reveal significant accuracy drops: 10% for modality classification arriving to and accuracy of $86.60\% \pm 3.36\%$ and 17% for content classification reaching an accuracy of $63.19\% \pm 4.40\%$. These reductions highlight the critical role of temporal dynamics in all classification tasks, with a more significant impact on content, which is more dependent on temporal information due to their complexity.

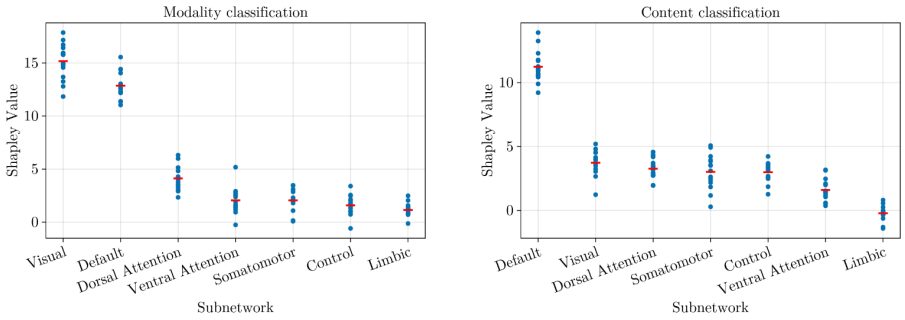

Fig. 1. Distribution of Shapley values for modality classification and content classification.

Figure 1 presents the contribution of Yeo 7 subnetworks computed with Shapley values for classifying narrative using the presented machine learning model. The distribution of Shapley values for each subnetwork is displayed, reflecting the results of retraining the model 15 times to ensure robustness against variations due to initial initialization. The red line represents the average of these 15 Shapley values. In the modality classification task, the visual subnetwork emerges as the most influential, followed by the default mode subnetwork (see Figure 1 (a)). This aligns with the intuitive notion that visual processing is essential for distinguishing between movies and audio stories. For the content classification task, the high value of the default mode subnetwork suggests its influence in understanding the meaning and content of the stimuli as suggested by previous studies that has highlighted the default mode subnetwork's involvement in higher-order cognitive functions, such as narrative comprehension [1, 7] (see Figure 1 (b)).

Acknowledgments

This work has been supported by the French government, through the UCA DS4H Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-17-EURE-0004. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

1. Christopher Baldassano, Uri Hasson, and Kenneth A. Norman. Representation of Real-World Event Schemas during Narrative Perception. *The Journal of Neuroscience*, 38(45):9689–9699, November 2018.
2. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777. Curran Associates Inc., 2017.
3. Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. "narratives", 2020.
4. Maria Giulia Preti, Thomas AW Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, 160:41–54, 2017. Functional Architecture of the Brain.
5. Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.
6. Lloyd S Shapley. Notes on the n-person game—ii: The value of an n-person game. 1951.
7. Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1):12141, July 2016.
8. Ann E. Sizemore and Danielle S. Bassett. Dynamic graph metrics: Tutorial, toolbox, and tale. *NeuroImage*, 180:417–427, 2018. Brain Connectivity Dynamics.
9. B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, September 2011.
10. Koene R. A. Van Dijk, Trey Hedden, Archana Venkataraman, Karleyton C. Evans, Sara W. Lazar, and Randy L. Buckner. Intrinsic Functional Connectivity As a Tool For Human Connectomics: Theory, Properties, and Optimization. *Journal of Neurophysiology*, 103(1):297–321, January 2010.
11. Roel M. Willems, Samuel A. Nastase, and Branka Milivojevic. Narratives for Neuroscience. *Trends in Neurosciences*, 43(5):271–273, May 2020.