ToM (Theory of Mind)-ML: Machine Learning predicts Mentalization

Varad Srivastava [1], Minaxi Goel [2]

[1] Indian Institute of Technology Delhi
[2] Magicbricks

Abstract

Introduction

Mentalizing is defined as "being engaged in a form of (mostly preconscious) imaginative mental activity that enables us to perceive and interpret human behavior in terms of intentional mental states," (Allen, Fonagy, and Bateman 2008) such as needs, desires, thoughts, feelings, intentions. In previous studies, impairment in mentalization has been linked to various psychological disorders such as autism (Frith 2001; Castelli et al. 2002; White et al. 2011; Abell, Happé, and Frith 2000), psychopathy (Decety et al. 2013), and schizophrenia (Russell et al. 2006). In light of these studies, we present an effort to model the prediction about mentalization from neural activity using predictive modeling.

Method

We used the task-evoked functional brain activity data from the Social Cognition (Theory of Mind) domain of the Human Connectome Project, where 339 participants were presented with animated videos of shapes (circles, triangles, or squares) either interacting with each other or moving randomly on the screen. These interacting shapes have shown evidence for mental state attributions (Castelli et al. 2000; Abell, Happé, and Frith 2000; White et al. 2011). This explains that an individual can perceive them, corroborating the mechanism of mentalizing. When PCA, a dimensionality reduction method, was applied on these neural recordings, we observed that the two classes are linearly separable in this space, with the exception of few scans (Figure 1). We used logistic regression to model the prediction of mentalization (i.e., if an individual is able to infer complex mental states in the interacting shapes, like being involved in persuading, bluffing, mocking, surprising one another or even depicting an intention to deceive) from both whole-brain activities as well as the activity in only the 36 parcels belonging to four ROIs, known to support ToM (temporo-parieto-occipital junction (TPOj), medial prefrontal cortex (mPFC), auditory association (AA) area, and lateral temporal cortex (LTC)). We also checked for overfitting by using L2 regularization, and evaluating the model in two ways: splitting the data 70-30 into training and testing sets and using repeated k-fold cross-validation for 10 folds and 3 repeats.

Results and Discussion

The machine learning model was observed to be 95.58% accurate in making predictions about mentalization on the testing set (Table 1 A). Interestingly, the model trained on the activity in only the 36 parcels (belonging to the four ROIs) achieved an accuracy of 92.64% (Table 1 B). To further validate our results, we plotted the confusion matrices for both classifications (Figure 2). The confusion matrices depict that there were 9 misclassifications in the first case (Figure 2 (a)), and neural activities in only 15 instances (out of 203) were misclassified in the second case (Figure 2(b)). In addition to this, repeated k-fold cross-validation returned 94.6%

mean accuracy (0.03 SD) for the model trained on the activity in 360 parcels, i.e., all regions, and 90.5% mean accuracy (0.032 SD) for the model trained on the activity in 36 parcels, i.e., the four ROIs. We observe that mentalization can be reliably predicted on the basis of neural activity from these four regions alone. Therefore, we hypothesize that a predictive modeling approach similar to our model can be performed on data involving clinical population for diagnosis of disorders linked to impairment of mentalization (Kazeminejad and Sotero 2019), as well as for assessing the Mentalization-Based Treatments (Allen, Fonagy, and Bateman 2008), which is subject to further investigation and is within the future scope of this study.

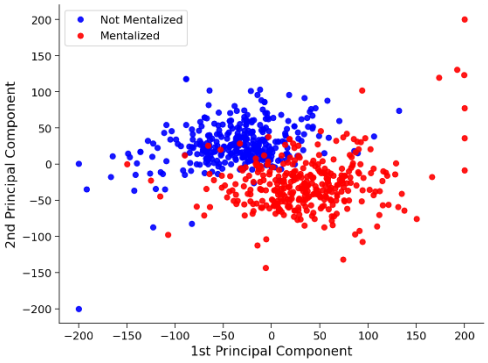

Figure 1: PCA performed on fMRI data from 4 ROIs. Each point represents scan of single participant.

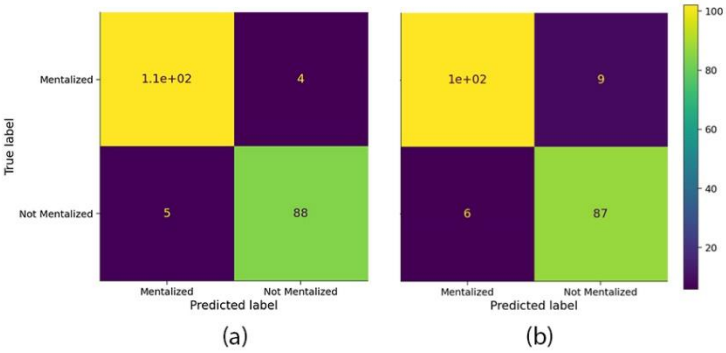

Figure 2: Confusion matrices for testing set of activity from (a) all regions (b) 4 ROIs

	Precision	Recall	f1-score	Support
A. All Regions				
Mentalized	0.96	0.96	0.96	111
Not Mentalized	0.96	0.95	0.95	93
Accuracy	0.96			
B. Four ROIs				
Mentalized	0.94	0.92	0.93	111
Not Mentalized	0.91	0.94	0.92	93
Accuracy	0.92			

Table 1: Classification metrics obtained when model trained on the activity in A: All Regions; B: Four ROIs

References:

Abell, F., F. Happé, and U. Frith. 2000. "Do Triangles Play Tricks? Attribution of Mental States to Animated Shapes in Normal and Abnormal Development." *Cognitive Development* 15 (1). https://doi.org/10.1016/S0885-2014(00)00014-9.

Allen, Jon, Peter Fonagy, and Anthony Bateman. 2008. "Mentalizing in Clinical Practice." *(2008)*.

Castelli, Fulvia, Chris Frith, Francesca Happé, and Uta Frith. 2002. "Autism, Asperger Syndrome and Brain Mechanisms for the Attribution of Mental States to Animated Shapes." *Brain* 125 (8). https://doi.org/10.1093/brain/awf189.

Castelli, Fulvia, Francesca Happé, Uta Frith, and Chris Frith. 2000. "Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns." *NeuroImage* 12 (3). https://doi.org/10.1006/nimg.2000.0612.

Decety, Jean, Chenyi Chen, Carla Harenski, and Kent A. Kiehl. 2013. "An FMRI Study of Affective Perspective Taking in Individuals with Psychopathy: Imagining Another in Pain Does Not Evoke Empathy." *Frontiers in Human Neuroscience*, no. SEP. https://doi.org/10.3389/fnhum.2013.00489.

Frith, Uta. 2001. "Mind Blindness and the Brain in Autism." *Neuron*. https://doi.org/10.1016/S0896-6273(01)00552-9.

Kazeminejad, Amirali, and Roberto C. Sotero. 2019. "Topological Properties of Resting-State FMRI Functional Networks Improve Machine Learning-Based Autism Classification." *Frontiers in Neuroscience* 13 (JAN). https://doi.org/10.3389/fnins.2018.01018.

Russell, Tamara A., Emanuelle Reynaud, Catherine Herba, Robin Morris, and Rhiannon Corcoran. 2006. "Do You See What I See? Interpretations of Intentional Movement in Schizophrenia." In *Schizophrenia Research*. Vol. 81. https://doi.org/10.1016/j.schres.2005.10.002.

White, Sarah J., Devorah Coniston, Rosannagh Rogers, and Uta Frith. 2011. "Developing the Frith-Happé Animations: A Quick and Objective Test of Theory of Mind for Adults with Autism." *Autism Research* 4 (2). https://doi.org/10.1002/aur.174.