Elucidating Cognitive Processes Using LSTMs

Pedro F. da Costa (pedro.ferreira_da_costa@kcl.ac.uk)
Centre for Brain and Cognitive Development, Birkbeck, UK
Centre for Neuroimaging Sciences, Institute of Psychiatry,
King's College London, UK

Sebastian Popescu (s.popescu16@imperial.ac.uk)
Computational Cognitive and Clinical Laboratory,
Imperial College London, UK

Robert Leech (r.leech@kcl.ac.uk)
Centre for Neuroimaging Sciences, Institute of Psychiatry,
King's College London, UK

Romy Lorenz (Romy.Lorenz@mrc-cbu.cam.ac.uk)
University of Cambridge, UK
Max Planck Institute for Human Cognitive and Brain Sciences, Germany.

Abstract

Despite several decades of functional neuroimaging research the relationship between brain networks and cognition remains elusive. This is because the taxonomy of cognitive processes was developed largely blind to the functional organization of the human brain. In this work, we leverage recent advances in artificial neural networks to gain insights into shared cognitive processes among six different cognitive tasks. We trained a single recurrent neural networks (RNN) to perform cognitive tasks. In this manner, we were able to evaluate shared representations between multiple cognitive tasks without relying on predefined cognitive processes. Next, we tested if the learned representations provide a good explanation for human brain activation patterns associated with these tasks. While we found little similarity between the RNN's learned representation and real brain data, our approach offers a roadmap to gain more mechanistic insights into how cognitive processes map to brain networks with potential important implications for studying cognitive dysfunction in disease.

Keywords: computational modelling; LSTMs; cognitive processes; human neuroimaging

Introduction

Understanding how brain networks give rise to cognitive processes is one of the primary endeavours of cognitive neuroscience. For this, cognitive neuroscientists instruct humans to perform cognitive tasks while undergoing functional neuroimaging. These cognitive tasks are designed to tap specific cognitive processes. However, the taxonomy of cognitive processes used in the field has largely been developed based on psychology theory; a discipline predating neuroimaging research (Bilder et al., 2009). Therefore it is unclear how well this taxonomy relates to the neurobiological underpinnings of cognitive processes. Recent empirical findings indicate the urgent need for a revised, data-driven (Eisenberg et al., 2019) and neurobiologically-informed cognitive taxonomy (Lorenz et al., 2018). Artificial Neural Networks (ANNs) have shown promising parallels with their neurobiological counterparts (Cueva and Wei, 2018; Carnevale et al., 2015) and, contrarily to brain systems, ANNs allow full access to its inner calculations. The understanding of the ANN's shared representations across tasks and its mechanisms might shed light on the relationship between cognitive paradigms and the structures that originate function (Kriegeskorte and Kievit, 2013).

The primary goal of this work was to create an ANN, that effectively solves six cognitive tasks, such as working memory, reaction-time and inhibitory control tasks, taking inspiration from the work developed by (Yang et al., 2019). A secondary aim consisted of trying to interpret the artificial network's representations and understand how specific inputs are modelled by the network. By studying the ANN's activations for each given task we were able to gain some insights into how each cognitive task was parameterized by the network. Finally, we also studied similarities in task representation between the ANN and large-scale brain networks when solving the same tasks.

Methods

Model architecture

We employed a RNN to solve the different tasks. In particular, we employed LSTMs, as they contain an additional state cell that can be subsequently analysed and avoids the vanishing gradient problem for large temporal dependencies (Hochreiter and Urgen Schmidhuber, 1997) . The entire model was trained end-to-end through supervised learning using batches containing an equal number of trials from every task trained. Stochastic Gradient Descent was applied for the optimization of the objective function. Most of the model's hyperparameters were chosen using empirical heuristic methodology. The

model was trained with a first hidden-layer of 256 LSTM units, a second hidden-layer of 128 LSTM units and an output layer with a softmax as the activation function, as the predicted response is categorical.

Cognitive tasks

To train the model, an artificial dataset was generated with 2000 trials simulating six tasks that the model was intended to learn: the Go Task, the Anti Go Task, the Memory Go Task, the Memory Anti Go Task, the Reaction Go Task, and the Reaction Anti Go Task. An example of a Go task trial is depicted in Figure 1. The input signal that is fed into the model can be decomposed into 4 different aspects – the rule signal, the fixation unit and the two modalities. The rule signal corresponds to a one-hot vector which determines which one of the six tasks is being executed. The fixation unit cues when and whether the model should react to the stimulus or not. The latter 64 nodes represent two separate modalities of 32 nodes each. Each modality can display an arbitrary signal, by activating one of its 32 nodes, which will define the expected response from the model. To simulate noise in underlying neural representations, the input matrix is combined with additive Gaussian noise (i.e. *N*(0;0.1)). In total there are $N_{in} = 1+6+32*2 = 71$ input units. Each one of them runs for 600 iterations, which accounts for one trial of a given task. The output signal (33 x 600) is a one-hot vector representation of the desired action of the model. It is composed of a fixation unit and a 32-unit reaction output, that represents the expected reaction of the network based on the stimulus direction. In total there are $N_{out} = 1+32 = 33$ output units. This trial structure is common to every task to allow consistency for the network's input and output. The task being resolved dictates the moment and the direction of the reaction to the stimulus.

Comparison with Meta-analytic Terms

To investigate whether the ANN's learned representation is consistent with brain activation associated with these tasks, the online platform for large-scale meta-analysis of neuroimaging data, Neurosynth (Yarkoni et al., 2011), was used. Neurosynth allowed us to identify meta-analytic brain activation maps based on specific terms that most closely related to the six tasks learned by the ANN: (1) "finger tapping", "motor task"; (2) "reaction time", "reaction times"; (3) "switch", "switching"; (4) "wm task", "working memory"; (5) "nogo" and (6) "incongruent". Next, we extracted meta-analytic brain activation within seven large-scale brain networks (Yeo et al., 2011) and cross-correlated these network-based activation maps, resulting in seven neural representational similarity matrices that could be compared to the ANN's similarity matrix. The ANN's similarity matrix was built on the correlations of variance across tasks.

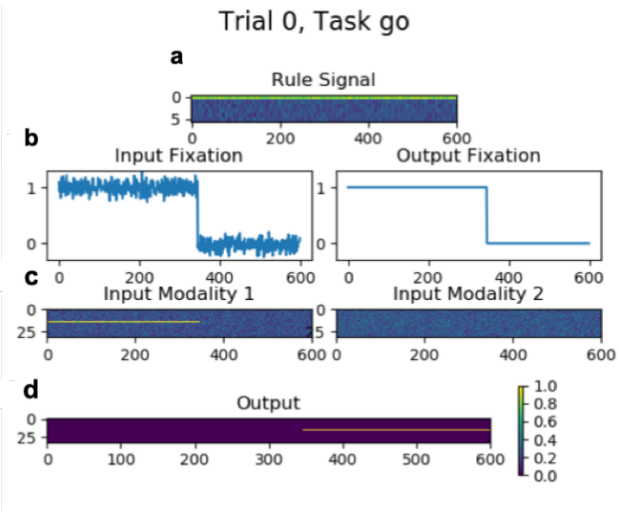

Figure 1: Example of a Go Task trial. a: Rule signal indicates the task being considered. b: Fixation unit. It falls to zero when the model is expected to react. c: Expected output of the fixation unit. The model is expected to mimic the fixation unit, in order to understand when it should react. d: Input modalities which simulate the direction of the stimulus. e: Expected output signal.

Results

The first experimental objective was to successfully train a model to simultaneously perform 6 different cognitive tasks. This objective was achieved with a general accuracy of 93% but , comparatively, the model underperforms on the Reaction Time tasks where it only achieves 66% accuracy. It is also notable how, for the other 4 tasks, the accuracies on the 200 trials are alike, with special emphasis on the Go and Anti tasks (i.e. 85.9% and 85.8% respectively). When the LSTM layers were substituted by Dense layers or vanilla-RNN layers, the model got stuck on local minima with around 50% accuracy.

Model Analysis

To characterise how each layer contributed to the model performance, the different layers' activations were analysed with the non-linear dimensionality reduction technique, t-SNE (Van Der Maaten and Hinton, 2008). Through this method, it is possible to analyse both the final outcome and the activations of the inner layers, including the cell states of the LSTM units. It is evident from Figure 2a that trials are organised by when the model has to react, with later responses occupying the distal region of each cluster and the earlier ones being represented in the proximal regions. It is also evident from Figure 2b, that there is a significant similarity between trials of a given task across the cell states of the

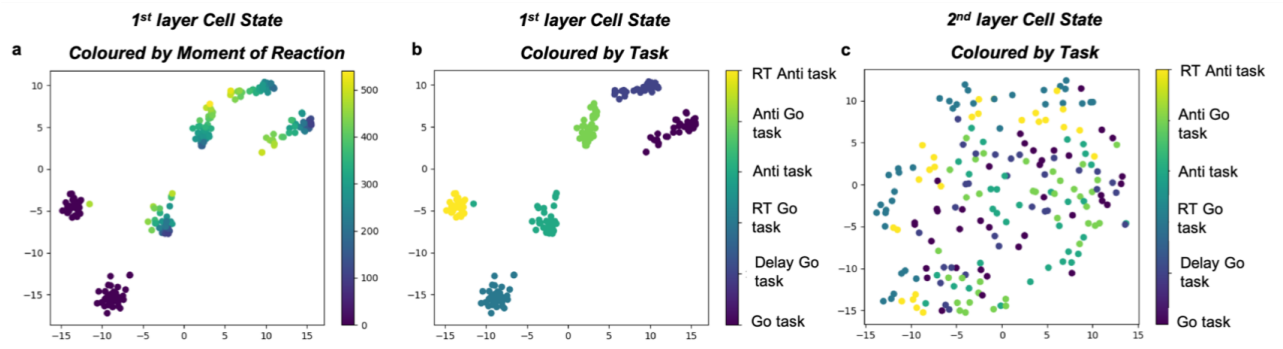

Figure 2: Four t-SNEs plots, each representing 200 trials as scattered points relating to the model's cell state. a: The activations gradually change for each cluster depending on the model's moment of reaction. b: There is a clear clustering of activations depending on the task being solved. c: The second hidden-layer activations has a more abstract signal and activation variance is high for each task.

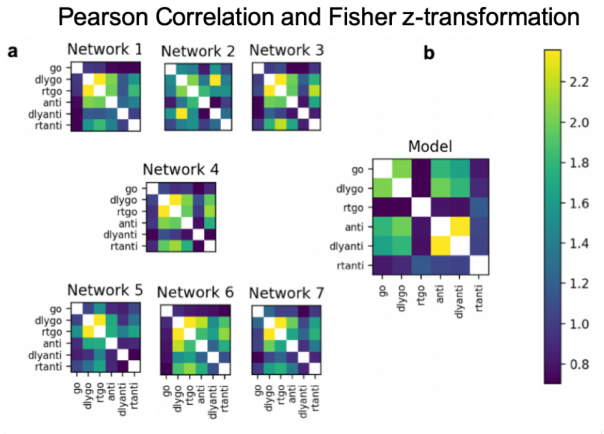

Figure 3: Similarity matrices for the Neurosynth meta-analytic maps on the 7 different Yeo networks using Pearson correlation followed by Fisher z-transformation (a) and for the ANN's variance across tasks (b). Colour codes correlation intensity.

first layer of the LSTM. The 6 clusters are grouped by the task being performed, which demonstrates how the model represents how different tasks relate to different expected reactions. Finally, the t-SNE analysis of the second hidden layer cell state, depicted in Figure 2c, does not present clustering of trials as the representations for a given task are more dissimilar. This is expected when considered that each second layer unit is affected by the activations of every first layer unit. Posteriourly, a variance study identified both Reaction-Time tasks with the larger amount of variance on the activation of its trials.

Blocking of neural unit responsible for inhibition

Unit 28 from the first hidden layer of the model displayed high variance across trial activations on several tasks, with special emphasis on the Go, Delay Go, Anti and Delay Anti tasks. To further understand its effects on the overall output of the model, the unit was eliminated by reducing its weights to zero. The model was subsequently analysed and it was observed that the model's accuracy dropped sharply to 49%. This effect was present across tasks with the exception of the Reaction tasks. The outputs were compared to the original model and it was observed that without unit 28 the model was not able to follow fixation before the expected reaction. The altered model reacts to the stimulus the moment it is shown, not being able to solve tasks that require inhibitory behaviour.

Exploring comparisons with brain activations

We considered whether and where the neural network's approach for modelling tasks would be similar to how the human brain processes similar types of tasks. To do so, it was first analysed how the model's activations correlated for different tasks. This similarity matrix between tasks could then be compared to an equivalent similarity matrix created from neural data taken from a large meta-analytic database of brain activation patterns (the Neurosynth database). Here, the 7 brain networks were considered separately.

There is no significant correlation between the model and the networks' similarity matrices after applying Fisher z-transformation (Figure 3). The low correlation of the Reaction Time tasks ("rtgo", "rtanti") with all other tasks in the model stands out as markedly different to the meta-analytic results. This result is by itself important as it suggests the model learned an efficient mechanism distinct from how the brain relates the different tasks.

Discussion

The ANN identified in this work efficiently solved different cognitive tasks simultaneously; however, it did so in a way that does not appear to correlate with how similar tasks are represented in the brain. This lack of similarity is interesting in itself; it suggests multiple solutions for mechanistically performing cognitive tasks can exist and the cognitive taxonomy underlying human cognition may be just one of the many possible cognitive taxonomies that could plausibly exist (Poldrack and Yarkoni, 2016). Despite the different mechanism displayed, unit 28 of the first hidden-layer seems to control the model's inhibition system or sustained attention. It constrains the model's reaction until it receives the instruction to follow the stimulus. In the human brain, similarly to our model, damages to the frontal lobes, caudate nucleus and subthalamic nucleus result in a lack of control of inhibition (Carnevale et al., 2015; Hampshire and Sharp, 2015).

Future work is needed to better refine the artificial network model, e.g., by adding in more biologically-plausible constraints, capturing a wider array of cognitive tasks, and training the model on more ecologically valid stimuli and tasks. The tasks don't correspond completely to the terms obtained from the meta-analysis. Instead of using the Neurosynth platform, humans could be tested using fMRI on the same tasks the model is executing. The model architecture/hyperparameters could also be explicitly optimized to maximize the overlap between network and neural structure across tasks (e.g., by using correlation), rather than optimizing model performance as we have done here.

There are limitations to our approach that justify the lack of correlation between the tasks of the ANN and the tasks from the meta-analysis. Still, there are several possibilities to bridge that gap. As better models, capturing a closer correspondence between artificial neural network and neural representations are developed, they will be useful for understanding the space of cognitive processes. In this sense, this work serves as a roadmap for a better insight into the mechanisms of cognitive processes.

Acknowledgements

Pedro F. da Costa is supported by the Marie-Sklodowska Curie Action Training Network. Robert Leech is supported by an MRC Project Grant. Romy Lorenz received support by the EPSRC (P70597) and is funded by the Wellcome Trust (209139/Z/17/Z).

References

Bilder, R. M., Sabb, F., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., Freimer, N. B., and Poldrack, R. A. (2009). NIH Public Access. 14(4).

Carnevale, F., DeLafuente, V., Romo, R., Barak, O., and Parga, N. (2015). Dynamic Control of Response Criterion in Premotor Cortex during Perceptual Detection under Temporal Uncertainty. *Neuron*, 86(4):1067–1077.

Cueva, C. J. and Wei, X.-X. (2018). Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *ICLR 2018 Conf.*, pages 1–19.

Eisenberg, I., Bissett, P., Enkavi, A. Z., Li, J., MacKinnon, D., Marsch, L., and Poldrack, R. (2019). Uncovering mental and neural structure through data-driven ontology discovery.

Hampshire, A. and Sharp, D. (2015). Inferior PFC Subregions Have Broad Cognitive Roles. *Trends Cogn. Sci.*

Hochreiter, S. and Urgen Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain.

Lorenz, R., Violante, I. R., Monti, R. P., Montana, G., Hampshire, A., and Leech, R. (2018). Dissociating frontoparietal brain networks with neuroadaptive Bayesian optimization. *Nat. Commun.*

Poldrack, R. A. and Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure.

Van Der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.*, 9:2579–2605.

Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T., and Wang, X.-J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.*

Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D., and Wager, T. (2011). NeuroSynth: a new platform for large-scale automated synthesis of human functional neuroimaging data. In *Front. Neuroinformatics Conf. Abstr. 4th INCF Congr. Neuroinformatics*.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.*