Distinct Computational Models of Reading Correspond to Distinct but Similar Neural Activation Patterns

William W. Graves (william.graves@rutgers.edu)
Department of Psychology, Rutgers University – Newark, 101 Warren Street
Newark, NJ 07102 USA

Amulya Bidar Nataraj (ab1670@scarletmail.rutgers.edu)
Information Technology and Analytics, Rutgers Business School, 1 Washington Park
Newark, NJ 07102 USA

Abstract:

A long-standing debate in cognitive computational neuroscience relates to the merits of modeling cognition using brain-inspired artificial neural networks with distributed representations compared to models using symbolic representations. Using established examples of each type of model from the domain of reading, we directly evaluated each with respect to its ability to capture correspondences among stimuli that matched with neural correspondences among the stimuli. Specifically, we focused on the internal feature representations situated between model inputs and outputs in a feed-forward distributed model of reading and in the symbolic dual-route model of reading. Word representations from the models were vectorized and pair-wise correlations were calculated among 464 words. To test for brain areas where activation-based word representations correlated with model-based representations, a searchlight analysis was performed across the whole left hemisphere cortex, and specifically within an atlas-defined region of interest in the fusiform gyrus. Both models showed similar correspondence with activation in anterior lateral temporal and fusiform regions, while only the distributed model correlated with activation in the inferior frontal gyrus language-related cortex. Overall, these results suggest that both modeling approaches capture neurally relevant information. The distributed model in particular, however, may capture more task-relevant information for reading aloud.

Keywords: reading; representation; symbolic; networks; brain

Introduction

A fundamental debate in computational approaches to cognition is whether the most useful approach to modeling involves distributed or symbolic representations (Bowers, 2017). Computational models of each type are particularly well-characterized in the domain of reading. Distributed, artificial neural network (ANN) models of reading consist of neurally inspired units connected by weights. Rather than starting with representations for whole words or pronunciation rules, the solution for mapping between letter strings and sounds develops by automatically tuning the weights using machine learning algorithms (Plaut, McClelland, Seidenberg, & Patterson, 1996). The dual-route cascaded (DRC) model of reading, however, consists of explicitly coded, symbolic representations for rules that specify mappings between letters and sounds. Exceptions to these rules are handled primarily by a separate route consisting of orthographic (visual word form) and phonological (sound word form) lexicons. The DRC model also contains weighted connections, but these connections are hand-tuned rather than learned (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). While these models have been extensively evaluated in terms of behavioral performance data (Coltheart et al., 2001; Plaut et al., 1996), evaluation in terms of neural data has so far only been indirect.

Here we directly compare ANN and DRC models of reading in terms of their ability to fit neural data from humans reading words aloud. The neurally inspired nature of the distributed ANN model presumably makes it a better candidate than the symbolic DRC model for corresponding with neural data. Alternatively, these models may simply offer different levels of description. For example, an ANN model based on combining distributed features might show neural correspondence with areas that either represent those features directly or with convergence zones that guide the coordination of those features (Damasio, 1989). The use of discrete, symbolic representations in the DRC model, on the other hand, is analogous to findings of discrete neural coding for high-level concepts such as person identity (Quian Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Therefore, we expect similar neural correspondence between the models to the extent that they both capture important similarities among words. We expect differences, however, in areas that may be coordinating features rather than representing discrete high-level units such as words or rules. Specifically, we tested for correspondence between model representations and brain representations within the left hemisphere, in order to focus on language-related regions, and the FG in particular, due to its consistent association with reading (McCandliss, Cohen, & Dehaene, 2003).

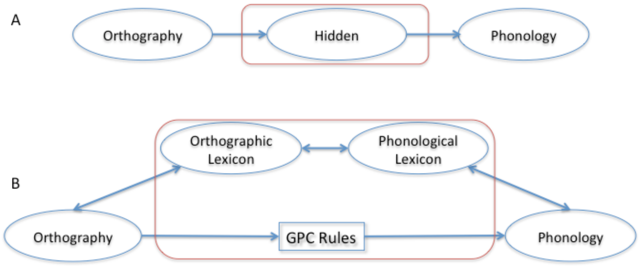

Figure 1: Schematic of the ANN (A) and DRC (B) models of reading. The red outline indicates areas from which tested representations were derived.

Methods

The ANN model was a re-implementation of the feed-forward model of reading by Plaut et al. (1996), but with a novel coding of orthographic inputs such that letter identity was coded categorically while letter position was coded probabilistically (Graves, 2017). This model was trained to pronounce 2998 monosyllabic words, 464 of which were later read by humans during fMRI. We specifically focused on the hidden unit representations, situated between orthographic inputs and phonological outputs in the three-layer model. The learned input-output solution is stored as connections to and from the hidden layer. This layer has 100 units, so the model-derived representation for each of the 464 test words consisted of a 100-unit vector for each word. 20 versions of this model, identical except for re-initialization of the weights, were run. Stimulus-stimulus correlation matrices of hidden unit values were generated for each model and averaged.

For the DRC model we used a downloadable version (http://www.cogsci.mq.edu.au/~ssaunder/DRC/). By analogy with the ANN model, we derived representations for each word from the orthographic and phonological lexicons, and the grapheme-phoneme-correspondence (GPC) rules. The lexical representation for each word consisted of the final model activation value for the word, and a 0 for the other 2997 words in the full set. The final activation value for the GPC route was also applied to each of the activated GPC rules for each word, from a set of 2033 total rules. Combining across both lexicons and the rules, each word consisted of a vector of 8029 units (2998 * 2 because there are two lexicons, plus 2033 rules), where the vast majority of slots had a value of 0. Both models are shown schematically in Figure 1.

Functional MRI data were obtained from 18 participants reading aloud 464 words, as described previously (Graves, Desai, Humphries, Seidenberg, & Binder, 2010). The fMRI data from the rapid event-related design were deconvolved into discrete neural events using least squares sum regression (Mumford, Turner, Ashby, & Poldrack, 2012). Model representations were compared to cortical activation using Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). RSA was applied using a series of searchlights throughout the left hemisphere, and in the atlas-defined fusiform gyrus (FG) region, using PyMVPA (Hanke et al., 2009). This yields a 2nd-order Spearman correlation for each model-based and brain-based searchlight tested. Coefficients were smoothed with a 6 mm isotropic kernel. Voxels in the left-hemisphere cortex were thresholded at $p < 0.001$, with an extent threshold of 221 μl (corrected $p < 0.05$). Voxels restricted to the FG were thresholded at $p < 0.005$, and extent threshold of 200 μl (corrected $p < 0.05$).

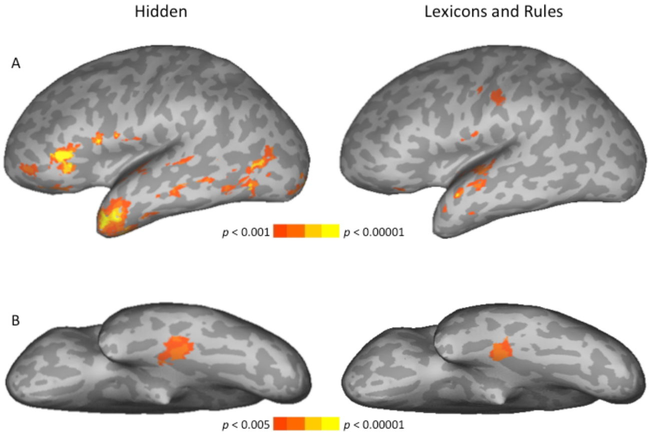

Figure 2: Results from whole left-hemisphere (A) and FG region (B) analyses for the ANN (left column) and the DRC (right column) models.

Results

Searchlight results for the left hemisphere are shown in Figure 2A, and the FG results are in Figure 2B. For the hemisphere analysis, the hidden layer of the ANN model showed correspondence with neural activation patterns primarily in anterior temporal cortex and inferior frontal gyrus (IFG), with a smaller cluster in the lateral occipito-temporal cortex (OT). The internal representations for the DRC model also showed correspondence with activation in the anterior temporal lobe, along with sensory cortex in the postcentral gyrus. No activation for the DRC model was seen in the IFG or OT. For the analysis focused on the FG, both models showed correspondence with activation in the anterior FG, with DRC activation slightly anterior to that of the ANN model. Direct contrasts showed no significant differences between activations associated with the two models.

Discussion

This is the first study we are aware of to directly compare distributed ANN and symbolic DRC models of reading in terms of neural data. Both models showed correspondence with fMRI data in areas of the anterior temporal and FG cortices. Only the ANN was associated with activation in the IFG and OT. These results suggest that both distributed and symbolic models of high-level cognition capture correspondences among stimuli that are relevant to neural processing. We suggest the location of these correspondences in the IFG for the ANN model indicates involvement of convergence zones for coordinating the relevant stimulus features for reading aloud. At the same time, the lack of significant differences in a direct contrast, combined with similar activations in anterior temporal and FG areas, suggests that the models are capturing similar high-order correspondences among the stimuli, despite their very different architectures and representations. Future work extending the ANN to include more than one hidden layer will allow for even richer comparisons between model and brain representations.

Acknowledgments

We thank Dr. Blair C. Armstrong for providing a version of the LENS simulator code that could be compiled with later releases of Linux. This work was supported in part by NIH grant R00 HD065839.

References

Bowers, J. S. (2017). Parallel distributed processing theory in the age of deep networks. *Trends in Cognitive Sciences, 21*(12), 950-961.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*(1-2), 25-62.

Graves, W. W. (2017). Using representations from artificial neural network models of reading to reveal neural activation patterns for different reading computations. *Conference on Cognitive Computational Neuroscience 2017*, Archived at https://ccneuro.org/2017/abstracts/abstract_3000172.pdf.

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral Cortex, 20*, 1799-1815. doi:10.1093/cercor/bhp245.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics, 7*, 37-53.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis -- connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, Article 4. doi:10.3389/neuro.06.004.2008.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences, 7*(7), 293-299.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage, 59*, 2636-2643.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435*, 1102-1107.