Using single unit recordings in PDP and localist models to better understand how knowledge is coded in the cortex

Jeffrey S. Bowers (j.bowers@bristol.ac.uk)
Department of Experimental Psychology, 12a Priory Road
Bristol, BS8-1TU UK

Keywords: localist representation; distributed representation; grandmother cell; neural network

Introduction

There is long history of studies documenting that some neurons respond to images of objects, faces, and scenes in a highly selective manner. This includes neurons in the human hippocampus (e.g., the famous example of a neuron responding to images of the actress Jennifer Aniston) and neurons in high-level visual cortex in monkey (for reviews see Bowers, 2009; Ison, Quian Quiroga, & Fried, 2015). These findings have led to a growing interest in the claim that some neurons code for information in a localist ('grandmother cell') manner, as reflected in the many contributions to a recent special issue on this topic in the journal *Language, Cognition, & Neuroscience* (Bowers, 2017).

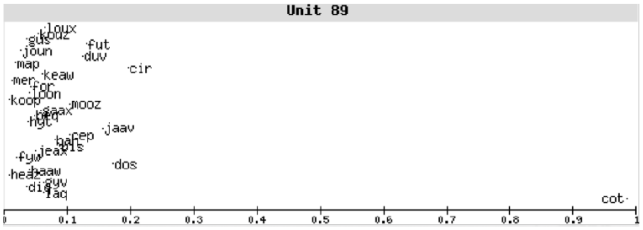
By contrast, it is only recently that interest in characterizing the selectivity of single units in connectionist networks has gathered speed. Critically, these studies also show that networks learn highly selective representations under a number of conditions, as detailed below. In this talk I will summarize recent research in my lab that explores the conditions in which artificial networks learn selective codes, and research comparing the responses of selective neurons and localist representations used in cognitive models. These findings suggest when and why some neurons in cortex respond in a highly selective manner, and highlight the biological plausibility of localist models in psychology.

Selective codes as a solution to the superposition catastrophe

In Bowers, Vankov, Damian, and Davis (2014, 2016) we carried out single-unit recordings on networks trained to co-activate multiple words at the same time in short-term memory (STM). We adapted models by Botvinick and Plaut (2006) who demonstrated that recurrent PDP networks can support human-like performance on STM tasks, and claimed that the models succeeded on the basis of co-activating learned distributed representations. This claim is important because it challenges the hypothesis that overlapping distributed representations result in blend patterns that are ambiguous, the so-called *superposition catastrophe* (Von Der Malsburg, 1986). The superposition catastrophe has been one of the key arguments in support of localist representations (Bowers, 2002; Page, 2000)

However, we showed that the Botvinick and Plaut (2006) and related models solved the superposition catastrophe by learning localist representations. Adapting an analytical tool developed by Berkeley, Dawson, Medler, Schopflocher, and Hornsby (1995), we carried out single unit recordings of the hidden units of trained networks. We showed that the models learned more selective codes when the superposition constraint became more challenging. For example, Figure 1 depicts a hidden unit (unit 89 of 200 hidden units) that responded selectively to the trained word 'cot' (taken from Bowers et al., 2014).

Figure 1



Furthermore, we found that recurrent networks of STM were only able to recall lists of novel words when they learned localist representations (Bowers et al., 2016), contrary to the widespread assumption that distributed codes are better able to support generalization. These findings extend our understanding of when and why some neurons respond selectively: Just as neurons in the hippocampus are thought to code information in a selective manner in order to support fast learning without forgetting (Marr, 1971), our findings suggest some neurons in cortex learn selective codes for the sake of STM.

Selective codes as a solution to some forms of arbitrary input-output mappings.

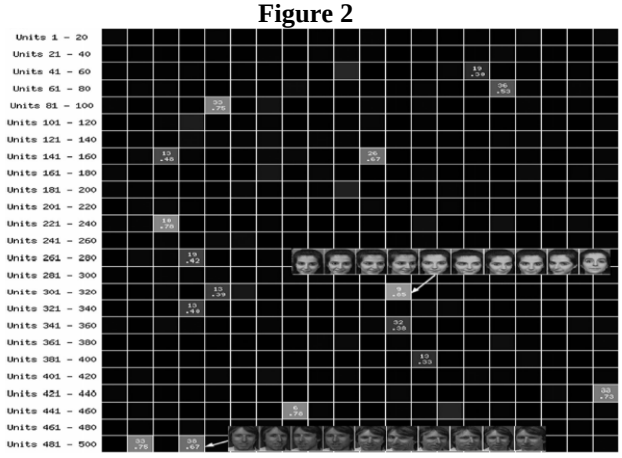
Recently there has been an explosion of interest in characterizing the selectivity of single hidden units in so-called 'deep' networks that achieve state-of-the-art performance on a range of tasks, including object and spoken word identification (for review, see Bowers, 2017). The striking finding is that these networks often learn highly selective representations even when trained on items one-at-a-time. This raises the question as to why we found that networks learned non-selective representations when trained on items one-at-a-time (Bowers et al., 2016).

Vankov and Bowers (2017) began to explore the conditions in which PDP networks learn selective and non-selective codes when trained on words one-at-a-time. Models learned non-selective distributed codes under a range of conditions, including when trained on many arbitrary input-output mappings. For example, a 3-layered model trained to map random patterns of binary inputs (with input units taking on an activation of 1 or 0) to another

random pattern of binary output units learned distributed codes.

However, we found one condition in which a 3-layered network learned localist codes: when trained on images of faces when input units took on continuous values and the model was trained on many-to-one mappings (with multiple different images of a given person mapping onto the same output representation). For example, Figure 2 depicts 16 localist units (units that are highlighted) out of 500 hidden units that selectively fire to a given face. To illustrate, the face images that activate units 312 and 404 are displayed.

Figure 2



We are currently carrying out more simulations to better understand the conditions in which networks learn distributed and localist codes when trained on items one-at-a-time. For example, is it many-to-one mappings that is critical, or the nature of the images themselves?

Comparing the selectivity of single neurons to the selectivity of single units in localist models in psychology.

Even when neurons are identified that selectively respond to images of one person or object within an experiment, it is often claimed that the neuron would responds to other (untested) categories of images. For example, Waydo et al. (2006) estimated that the most selective neurons observed in Quian Quiroga et al. (2005) study would respond to between 50-150 different people or objects if researchers had more time to find the relevant images. This is taken as inconsistent with grandmother cells.

However, Gubian, Davis, Alderman, and Bowers (2017) showed that the analysis of Waydo et al. (2006) is consistent with localist models in psychology. We carried out single-unit recordings in the Spatial Coding Model of visual word identification that represents ~30,000 words in a localist manner (Davis, 2010). Under parameter conditions that allow the model to correctly identify words we found that that the localist representations responded to approximately to 50 different words (e.g., the word DOG responds most strongly to the input DOG, but also responds above baseline to LOG, FOG, JOG, etc.). Page (2017) also provides evidence that localist models can account for single-cell recording data taken to support distributed coding. Together, these results highlight the computational reasons why some neurons in cortex respond in a highly selective manner, and show that localist (grandmother cell) representations are in fact consistent with single-cell recording data.

References

Berkeley, I. S. N., et al. (1995). Density plots of hidden unit activations reveal interpretable bands. Connection Science, 7, 167–186

Bowers, J. S. (2002). Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand. Cognitive Psychology, 45, 413–445.

Bowers, J.S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116*, 220-251.

Bowers, J. S. (2017). Grandmother cells and localist representations: A review of current thinking. *Language, Cognition, & Neuroscience, 32*, 257-273.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological Review, 121*, 248-261.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2016). Why do some neurons in cortex respond to information in a selective manner? Insights from artificial neural networks. *Cognition, 148*, 47-63.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*, 713-758.

Gubian, M., Davis, C.I., Adelman, J.S., & Bowers, J.S. (2017). Comparing single-unit recordings taken from a localist model to single-cell recording data: A good match. *Language, Cognition, & Neuroscience, 32*, 380-391.

Ison, M. J., Quian Quiroga, R., & Fried, I. (2015). Rapid encoding of new memories by individual neurons in the human brain. *Neuron, 87(1)*, 220-230.

Marr, D. (1971). Simple memory: A theory for archicortex. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 23–81

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. Behavioral and Brain Sciences, 23, 443–512.

Page, M. (2017). Localist models are compatible with information measures, sparseness indices, and complementary-learning systems in the brain. Language, Cognition and Neuroscience, 32(3), 366-379.

Waydo, S et al. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience, 26*, 10232–10234.

Vankov, I. I., Bowers, J. S. (2017). Do arbitrary input-output mappings in parallel distributed processing networks require localist coding? *Language, Cognition, & Neuroscience, 32*, 392-399.

Von Der Malsburg, C. (1986). Am I thinking assemblies? *In Brain theory* (pp. 161-176). Springer Berlin Heidelberg.