A Neural Model of Human Image Categorization

Eric Hunsberger (ehunsberger@uwaterloo.ca)
Peter Blouw (pblouw@uwaterloo.ca)
James Bergstra (jbergstra@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience
University of Waterloo, Waterloo, ON, Canada N2L 3G1

Abstract

Although studies of categorization have been a staple of psychological research for decades, there continues to be substantial disagreement about how unique classes of objects are represented in the brain. We present a neural architecture for categorizing visual stimuli based on the Neural Engineering Framework and the manipulation of *semantic pointers*. The model accounts for how the visual system computes semantic representations from raw images, and how those representations are then manipulated to produce category judgments. All computations of the model are carried out in simulated spiking neurons. We demonstrate that the model matches human performance on two seminal behavioural studies of image-based concept acquisition: Posner and Keele (1968) and Regehr and Brooks (1993).

Keywords: category representation; image categorization; neural engineering framework; vector symbolic architecture

Introduction

Although studies of categorization have been a staple of psychological research for decades, there continues to be substantial disagreement about how the mind represents information about unique classes of objects. Theories involving prototypes, exemplars, and explanatory schemas have all been shown to account for only a subset of known categorization phenomena, and progress toward a unified theory of category representation has been limited (for reviews, see Murphy, 2002; Machery, 2009; Smith & Medin, 1981). Historically, the difficulty in modelling category representation has been to balance generality and accuracy.

On one hand, many of the models developed from these theories have a fairly narrow scope of application. Consider, for instance, similarity-based accounts of concept reference; these models produce impressive results at matching human behaviour in tasks that involve feature comparisons (see Smith & Medin, 1981), but they do not generalize well to other tasks that require the use of deeper category knowledge or explanatory inferences (see Murphy, 2002).

On the other hand, approaches with greater scope tend to pay a price in terms of predictive accuracy or viability. For example, Barsalou's (1999) theory of perceptual symbol systems is a more or less unified account of category representation, but it lacks a corresponding computational model (Dennett & Viger, 1999). Rogers and McClelland's (2004) account of semantic cognition provides a powerful model that performs well across a range of categorization tasks, but employs both an idealized neural architecture and an idealized set of inputs (i.e. it is an abstract connectionist network that does not use raw percepts as input). Many researchers now recognize that object perception and conceptual cognition are not distinct (Palmeri & Gauthier, 2004), making it important that models integrate both perception and cognition.

In this paper, we argue that advances in our understanding of the visual system and new principles for the design of neural architectures can be used to overcome many of the difficulties in providing a viable, neurally grounded, computational model of image categorization. We use the techniques of the Neural Engineering Framework (NEF) (Eliasmith & Anderson, 2003) to develop a model of category representation that connects retinal activity to high level cognitive judgments using a class of vector-symbolic representations called semantic pointers (Eliasmith et al., 2012). The model receives natural images as input, produces category judgments as output, and carries out intermediate processing in simulated neurons. The proposed model replicates human performance on two independent studies of human judgment in prototype-based and exemplar-based image categorization, with no changes to the model. Semantic pointer architectures have been shown to support several important cognitive capacities (e.g. Stewart & Eliasmith, 2011; Eliasmith et al., 2012). Our study extends this line of research, showing that semantic pointers computed by a plausible visual system model can be used to replicate human category judgments.

Model Description

We developed a model of human image categorization that consists of a feed-forward visual perception model (similar to Hinton & Salakhutdinov, 2006) driving a vector-symbolic associative memory (see Gayler, 2003; Plate, 2003). The model was first constructed using a rate approximation of the spiking leaky integrate-and-fire (LIF) neuron model for the visual system and explicit vector computations for the associative memory. The model was then implemented fully in spiking neurons using the principles of the NEF.

The visual system component of the model is a sequence of feed-forward neural populations that compresses high dimensional input images into comparatively low dimensional vectors, which we refer to as *semantic pointers*. The first population, analogous to the retina and lateral geniculate nucleus (LGN), is a rasterized image, as would be captured by a conventional digital camera. Like the retina, a camera adapts to global lighting conditions and provides an image with standard intensity levels. Our (small) LGN population corre-

sponds to a square 30×30 image region that is best compared to a small portion from the centre of the visual field. The second population, analogous to visual area V1, comprises 2500 neurons with local connectivity: each neuron responds to a randomly chosen 9×9 patch in LGN. Neurons in the third (V2), fourth (V4), and fifth (inferotemporal (IT) cortex) populations (with size 900, 400, and 225 respectively) are connected to all neurons in the previous population. The activation pattern in the fifth population (with latency similar to visual area IT) is the semantic pointer representing the image stimulus. Representations generated in this manner are stored in an associative memory as category exemplars (during training), and used to probe the associative memory to yield a category judgment during testing (see Figure 2).

Adaptation to Natural Images

A large fraction of neuron cells in visual area V1 are well modelled as luminance edge detectors (Hubel & Wiesel, 1968; DiCarlo, Zoccolan, & Rust, 2012). There is mounting evidence that visual system neurons behave as they do because they continuously adapt to statistical patterns in visual stimuli (Olshausen & Field, 1996; Hyvärinen, 2009). Computer vision systems inspired by principles of adaptation to unlabelled visual images are among the best-performing computer vision systems, and reproduce several phenomena discovered by electrode recordings (Lee, Ekanadham, & Ng, 2008; Le et al., 2012). One strategy for adaptation to unlabelled images is the *autoencoder* (Ackley, Hinton, & Sejnowski, 1985; Rumelhart, Hinton, & Williams, 1985), which was first applied to images by Cottrell, Munro, and Zipser (1987).

The connection weights of our visual system model were trained as a deep autoencoder, with an additional ℓ_1 penalty on the hidden node activation rates to model the energy cost of spiking and encourage sparser activation patterns. The objective function for one layer is given by

$$O = \frac{1}{K} \sum_{i,k} \left(x_i^{(k)} - y_i^{(k)} \right)^2 + \lambda \sum_j |q_j - \rho| \qquad (1)$$

where $x_i^{(k)}$ is the value of visual node i for example k, $y_i^{(k)}$ is the autoencoder's reconstruction of node i example k, q_j is a running average of the activation of hidden node j, and λ and ρ control the importance of sparsity and the desired sparsity level, respectively. Uniquely, our autoencoder used an LIF response function as the feature activation function.

The autoencoder was trained on random 30×30 natural image patches chosen from the first 10 images of the van Hateren Natural Image Database (van Hateren & van der Schaaf, 1998). with each patch normalized to zero mean and unit variance. We trained only on un-whitened images, which contain the full spectrum of spatial frequencies. We found that whitening was not required to extract Gabor-like filters from the statistics of the natural images (Figure 1), and was in fact undesirable since it removed some low-frequency features important for classification.

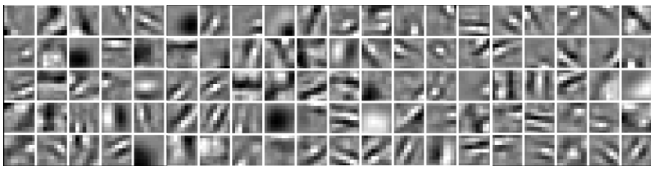

Figure 1: Filters from the first layer of the visual system, after autoencoder training on natural images. Like neurons in area V1, our model neurons detect luminance edges at a variety of frequencies and orientations.

Like Hinton and Salakhutdinov (2006), each layer of the autoencoder was pretrained individually; however, layers were pretrained as autoencoders, not restricted Boltzmann machines, allowing us to use an LIF response function for the neuron nonlinearity. The layers were then combined into a deep autoencoder and trained together using backpropagation.

Semantic Pointers: Memory and Retrieval

We refer to the vectors processed by the model as semantic pointers because they function as compressed symbol-like representations that encode the statistically relevant aspects of the information they represent (Eliasmith, in press). In the non-visual component of the architecture, semantic pointers representing the compressed images are bound with category labels using the mathematical operation of circular convolution (see e.g. Plate, 2003). Subsequently, the bound representations are added to the memory via superposition. This process is captured formally by Equation 2:

$$M = \sum_{i=1}^{N} (P_i * L_i) \qquad (2)$$

where P_i is a semantic pointer produced by the visual system from the i^{th} raw image, L_i is a vector representing the category label associated with the image, M is the memory pointer, and $*$ is the circular convolution operator.

Once the memory is built up with a number of learned category exemplars, it can be used to produce categorization judgments in response to novel input images via the use of an inverse of the convolution operation. This inverse operation probes the memory for the category label that is most likely to fit the input image on the basis of prior learning. As a whole, the categorization process conforms to the following mathematical description:

$$c = \text{argmax}_c \left[(P^{-1} * M) \cdot L_c \right] \qquad (3)$$

where c refers to the resulting category judgment, P^{-1} refers to the pseudoinverse of the semantic pointer corresponding to the test image, L_c refers to the label pointer of category c, and $a \cdot b$ refers to the dot product of a and b. For the rate model, these operations were implemented directly in vectors; for the spiking model, the operations were implemented in spiking LIF neurons using the NEF.

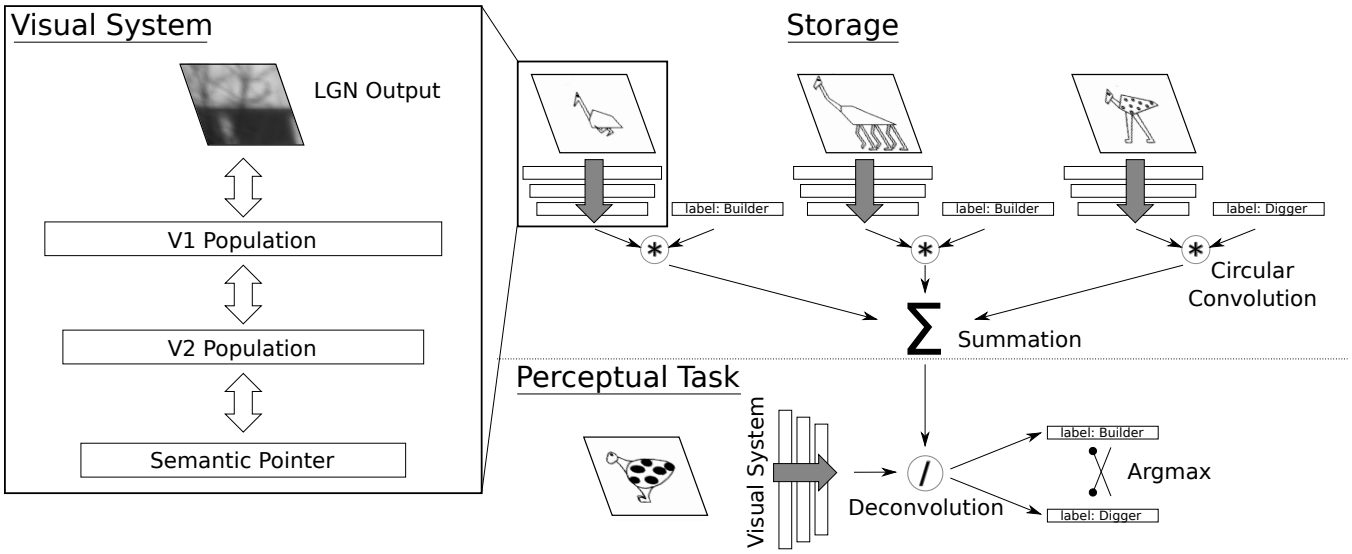

Figure 2: The schematic of our visual categorization model has three components. *Left:* The visual system comprises four populations of leaky integrate-and-fire neurons corresponding to the LGN, visual areas V1, V2, V4, and IT, which we take to represent a semantic pointer. The connections between these populations are adapted to natural scene statistics by unsupervised learning. *Upper Right:* The memory of our model is encoded as a single semantic pointer, which is the sum of several labelled training patterns (three are shown here). Labels have been bound to their corresponding image representations through the mathematical operation of circular convolution. *Lower Right:* At test time, our model labels visual stimuli by deconvolving the activity patterns of IT with the memory vector, and matching the result against several possible label decisions.

In short, the model builds category representations by storing compressed and labelled percepts, and produces categorization judgments by evaluating the similarity between an input percept and the exemplars stored in memory. However, since all labelled percepts are compressed into the same vector, there is significant interaction between stored percepts; this can be likened to creating a prototype based on the percepts. The model categorization system thus falls part way in between pure exemplar-based categorization and pure prototype-based categorization; it has elements of both.

Experiment 1: Prototype-based Categorization

To account for the sort of phenomena that have traditionally motivated prototype theories of category representation, we tested the model on a task from Posner & Keele's (1968) classic study of pattern classification. We chose to model Experiment 3 of the study, which was designed to test whether human subjects are learning about class prototypes when they only ever see distorted examples. In the study, subjects are trained to classify classify random dot patterns into three mutually exclusive categories. Each pattern consists of nine dots dispersed over a 30×30 grid, with each dot occupying one cell in the grid. The patterns used for training are generated from three prototypes; each training pattern is created by choosing a prototype pattern, and moving each dot according to a random distortion rule (see Figure 3.) Thirty (30) subjects were trained by corrective feedback to classify twelve 'high distortion' patterns (four from each category). After training, the subjects were asked to classify twenty-four patterns without feedback: patterns from the training phase (2 per prototype, 6 total), the prototype patterns (3), prototype patterns with a smaller degree of distortion (6), new highly distorted prototype patterns (6), and entirely random new patterns (3). Subjects were tested on these patterns on two consecutive days, in terms of both accuracy and reaction time.

The protocol for evaluating our categorization model was nearly identical. We presented the model with the twelve training images, and it stored the semantic pointers associated with the labels and the images into the model's memory (see Figure 2). Then we presented our model with each of the twenty-four testing patterns. Figure 4 compares the accuracy of our model to the classification accuracy of the human subjects. Since our model lacks motor output, we did not evaluate it on reaction time. Figure 4 shows the results of our model; in sum, the model performs much like the human subjects.

Experiment 2: Exemplar-based Categorization

To account for effects more commonly aligned with exemplar theories of category representation, we tested the model on a task from Regehr & Brook's (1993) study of the comparative influence of analytic and non-analytic processes on categorization behaviour. The study involves a number of experiments in which subjects are asked to classify simple drawings of imaginary animals into one of two categories. The animals all possess an analytic structure that varies along five binary dimensions (e.g. a round vs. angular body), but the exact perceptual manifestation of a particular dimension value (i.e.

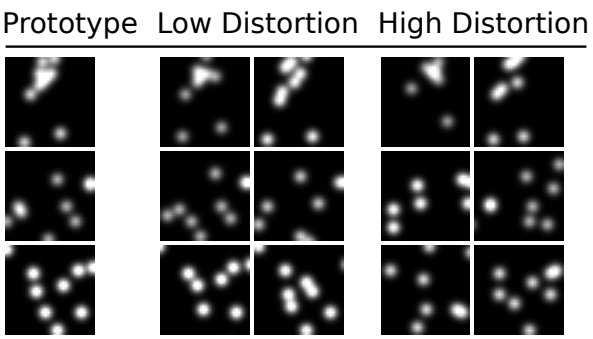

Figure 3: Sample stimuli for Experiment 1, modelling a classic study by Posner & Keele (1968). The dot patterns are created by distorting three randomly drawn prototype images (*left*) with low (*centre*) and high (*right*) levels of noise. Subjects are trained to classify a set of twelve high-distortion patterns and tested without feedback on the same prototypes at different distortion levels.

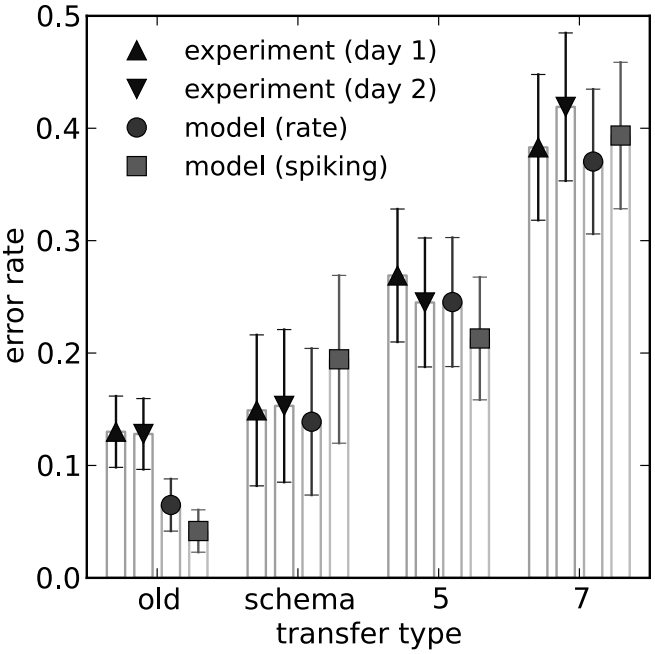

Figure 4: Comparison of human and model performance for Experiment 1. The model is able to account for human results when presented with the schema, low distortion (5), and high distortion (7) patterns. Occasional random errors by human subjects may explain the discrepancy on training examples. Error bars indicate 95% confidence intervals. Human data from Posner & Keele (1968).

feature) can vary across animals. For example, two animals might have round bodies, and thus be analytically equivalent to some extent, but the actual roundness of their respective bodies might be quite distinct (see Figure 5). This allows for the construction of stimuli sets that possess drawings that are analytically equivalent but perceptually dissimilar, along with drawings that are analytically distinct but perceptually similar. By training subjects through corrective feedback to classify these images into categories defined by an analytic rule, Regehr & Brooks were able to test hypotheses regarding the relative importance of perceptual similarity and analytic structure during categorization.

In the experiment 1C of Regehr & Brooks' study, 32 subjects are placed into one of two conditions and then trained to classify a set of eight images into two categories. For subjects in the first condition, the perceptual manifestations of a given dimension are constant across the images (See Figure 5, left). For subjects in the second condition, the perceptual manifestations of a given dimension vary across images (See Figure 5, right). Every subject was trained to learn one of four labelling rules based on analytic structure. The rules had the form: An image is a 'builder' if it has at least two of X, Y, and spots, otherwise it is a 'digger'. The criteria X and Y referred to things like "long neck", "angular body", "short legs" and so on (see Regehr & Brooks, 1993, for details). Training occurs through corrective feedback and is considered complete after five runs through the image set.

During the transfer phase of the experiment, subjects are asked to classify a set of sixteen images, eight of which are from the training set and eight of which are qualitatively similar, but new. The new images have been designed to pair up with a specific training image, and only differ on the dimension of "spots on body." Half of the new images belong in the same category as their twin from the training set, while the other new images have a different category from their twin. The idea motivating this experimental design is that if subjects attend primarily to the analytic structure of the images during testing, then they should make relatively few errors on the new bad transfer items (because both analytic structure and perceptual similarity favour the correct judgment). Alternatively, if subjects attend primarily to similarity to past exemplars, then they should make relatively more errors on the bad transfer items (because perceptual similarity and analytic structure favour opposing judgments). The study is designed to test the effect of appearance on subjects' use of structural vs. perceptual mental representations.

We model experiment 1C of Regehr & Brooks' study with the same model that we used in Experiment 1. We presented our model with the same eight training images used in the original experiment (though downsampled to fit in a 30×30 patch), drawn either in the composite style or in the individuated style. The semantic pointers created by the visual system, together with semantic pointers for the corresponding image labels, were stored into the model's memory, as described by Equation 2 shown in Figure 2. We tested the representations of our visual system by classifying the good-transfer and bad-transfer test images, as well as the original training images. The accuracy of our model in each case is presented in Figure 6. Our model provides a good match to human performance, and replicates the effect that perceptually individuated stimuli foster substantially different error profiles than perceptually un-individuated stimuli.

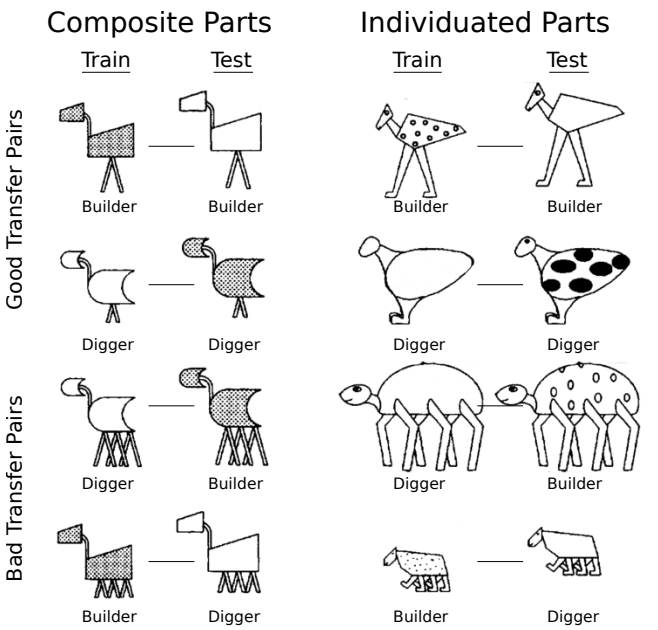

Figure 5: Sample stimuli for Experiment 2, modelling experiment 1C of Regehr and Brooks (1993). *(Left)* Images are composed of interchangeable (composite) feature manifestations. *(Right)* Images expressing the same attributes are drawn in a more coherent (individuated) style. Regehr & Brooks (1993) drew a distinction between *good transfer* and *bad transfer* test stimuli. A test stimulus is a good transfer case when the addition or removal of spots matches a training case with the same label, and a bad transfer case if adding or removing spots matches a training case with the opposite label. (Adapted from Regehr & Brooks (1993) Figure 2A).

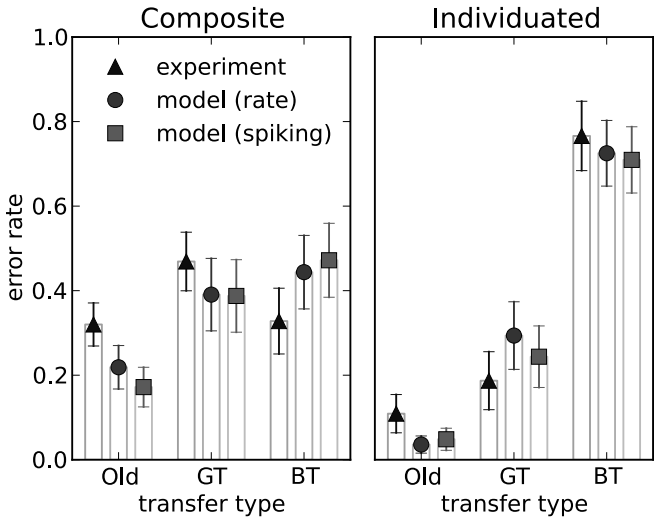

Figure 6: Comparison of human and model performance for Experiment 2. Our model accounts for the key difference in human performance on the good transfer (GT) versus bad transfer (BT) pairs for the individuated stimuli. Error bars indicate 95% confidence intervals. Human data from Regehr & Brooks (1993).

Discussion

Posner & Keele's (1968) study is considered to be seminal in the development of prototype theory, and the result that subjects categorize the training patterns and prototype patterns equally well is taken to indicate that the subjects are abstracting information about the prototypes during the training phase. Our model's replication of this performance provides good evidence that our approach is capable accounting for the sort of prototype effects that the study uncovered. Interestingly, the spiking version of the model performs slightly worse than humans on the prototypes, indicating that it might be performing a more exemplar-based classification. However, we hypothesize that adding more neurons to the associative memory will attenuate this effect.

Regehr & Brooks' (1993) study is more easily located in the tradition of exemplar theories of category representation. The fact that the model replicates the effects of interference from exemplar memories on more analytic categorization approaches suggests that it is well-equipped to deal exemplar-based phenomena. Moreover, the architecture of the model almost trivially assures that this is true—the contents of the associative memory essentially are exemplars produced from visual experience. It is thus reasonable to expect that phenomena found in studies using different kinds exemplars will be reproducible with the model.

Overall, the results of the simulations indicate that our model is able to account for an important set of phenomena closely associated with exemplar and prototype theories of category representation. The fact that the simulation employs a neural architecture for all stages of processing, and that it begins with raw image input, provides an important contribution to the current literature.

However, the model has several limitations as it stands. Nevertheless, we believe it is reasonable to expect that the architecture is capable of capturing an even wider range of phenomena. We identify two requirements of scaling that an architecture utilizing semantic pointers can likely achieve.

For one, it is possible to incorporate a more realistic account of category learning into the model. In the actual experiments, subjects learn the relevant categories through corrective feedback, and the feedback process continues either for a set number of trials, or until the subjects can accurately classify all of the items without error. By comparison, our model learns by memorizing a set of training images labelled with the correct category. In the model, the label/image relationships are forgotten when the model is shown another set of stimuli. However, the recent development of methods for introducing biologically plausible learning rules into the neural framework we employ indicates that this simplification could be removed in the future. Other cognitive models that make use of semantic pointers have already incorporated a form of reinforcement learning using such rules (e.g. Eliasmith et al., 2012).

Second, we believe it is possible to account for a broader range of categorization phenomena. The architectural prin-

ciples used in our model have also been used to construct what is currently the world's largest functional brain model, able to account for tasks involving serial-order recall, syntactic induction, and the manipulation of numerical concepts (Eliasmith et al., 2012). The fact that other large-scale cognitive models make use of the same representations and processes as this model provide good reason to think that a similar scale of functionality can achieved with models specifically focused on category representation. These two extensions will be the focus of future work.

Conclusion

This paper has presented a neural architecture for categorizing visual stimuli using a semantic pointer architecture. Our model replicates human behaviour on two important studies of visual categorization: Posner & Keele's (1968) and Regehr & Brooks' (1993). Modelling efforts have traditionally had to face the dilemma of choosing between plausibility and scope. The end-to-end neural model described here takes a suggestive first step in addressing this dilemma. Overall, this promise of scalability adds further theoretical significance to the empirical results we describe. The combination of a hierarchical visual model and a neurally implemented vector-symbolic architecture yields a new, effective approach to building models of category representation that are scalable, biologically plausible, and comprehensive, in that they capture the stages of processing from perception to judgment.

Acknowledgments

We gratefully acknowledge the financial support of NSERC, SSHRC, Canada Research Chairs, the Canadian Foundation for Innovation, the Ontario Innovation Trust, and the Banting Postdoctoral Fellowships Program.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science*, *9*(1), 147–169.

Barsalou, L. (1999). Perceptual symbols systems. *Behavioral and Brain Sciences*, *22*, 577-660.

Cottrell, G., Munro, P., & Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. In *Proc. COGSCI 9* (pp. 462–473).

Dennett, D., & Viger, C. (1999). Sort-of symbols? *Behavioral and Brain Sciences*, *22*(4), 613.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*, 415-34.

Eliasmith, C. (in press). *How to build a brain: A neural architecture for biological cognition*. Oxford, UK: Oxford University Press.

Eliasmith, C., & Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T., Choo, F.-X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202-1205.

Gayler, R. (2003). Vector symbolic architectures answer Jackendoffs challenges for cognitive neuroscience. In P. Slezak (Ed.), *ICCS/ASCS* (p. 133-138).

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, *313*(5786), 504–7.

Hubel, D., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Physiology*, *195*(1), 215-43.

Hyvärinen, A. (2009). Statistical models of natural images and cortical visual representation. *Topics in Cognitive Science*, *2*, 251–264.

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2012). Building high-level features using large scale unsupervised learning. In *Proc. ICML 29* (pp. 81–88).

Lee, H., Ekanadham, C., & Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. In *Proc. NIPS 20* (pp. 873–880).

Machery, E. (2009). *Doing without concepts*. Oxford, UK: Oxford University Press.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607-691.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature reviews. Neuroscience*, *5*(4), 291–303.

Plate, T. (2003). *Holographic reduced representation: Distributed representation for cognitive structure*. Stanford, CA: CSLI Publications.

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Experimental Psychology*, *77*, 653-663.

Regehr, G., & Brooks, J. (1993). Perceptual manifestations of analytic structure - the priority of holistic individuation. *Experimental Psychology: General*, *122*(1), 92-114.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rumelhart, D., Hinton, G., & Williams, R. (1985). *Learning internal representations by error propagation* (Tech. Rep. No. ICS-8506). UCSD Institute for Cognitive Science.

Smith, E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Stewart, T., & Eliasmith, C. (2011). Neural cognitive modeling: A biologically constrained spiking neuron model of the tower of hanoi task. In *Proc. COGSCI 33* (p. 656-661).

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Biological Sciences*, *265*(1394), 359-366.