Learning Context Sensitive Logical Inference in a Neurobiolobical Simulation[*]

Chris Eliasmith
Dept. of Philosophy and Dept. of Systems Design Engineering
University of Waterloo, Waterloo, Ontario

Introduction and model description

There remains a large difference between the kinds of models typical of cognitive neuroscience versus those typical of systems neuroscience: the former tend to be 'high-level', where components of the model are very large portions of cortex and the relevant behaviors are cognitive, whereas the latter tend to be 'low-level', where each component is a single cell and the relevant phenomena are sub-personal. This is true despite the fact that researchers in these areas share a similar interest in brain-based explanations of behavioral phenomena. In this paper I apply the neural engineering framework (NEF) described in Eliasmith & Anderson (2003) to describe a model that is both 'high-level' and 'low-level'. I do this by constructing a biologically detailed model of a traditionally cognitive phenonema – logical inference.

Logical inference, or deductive inference, is an ideal exemplar for this work because it is generally considered a phenomena that can only be explained with sophisticated, language-like processing. Indeed, it has been widely suggested that neurally plausible architectures do not naturally support structure sensitive computations, thereby demonstrating that understanding neural computation is not relevant for understanding cognitive function (Fodor & Pylyshyn 1988; Jackendoff 2002). However, since the early 1990s, there have been a series of suggestions as to how to incorporate structure sensitive processing in models employing distributed representations (including Spatter Codes (Kanerva 1994); Holographic Reduced Representations (HRRs; Plate); and Tensor Products (Smolensky 1990)). Some of these approaches have been used to build models of cognitive phenomena (Eliasmith & Thagard 2001).

However, none of these methods have been employed in a biologically plausible computational setting. The NEF provides principles by which various levels of description of neural function can be integrated. This model integrates the relevant physiological and anatomical data from frontal cortices (Wharton & Grafman 1998), HRR representations, and human performance on the Wason card selection task (Wason 1966; Cheng & Holyoak 1985). Specifically, I describe a dynamic spiking network model that learns the relevant transformations of structured representations in different contexts, based on past experience.

The network receives input from: a) VMPFC which provides familiarity of content information that is used to select the appropriate transformation (Adolphs *et al.* 1995); b) left language areas which provide HRR representations of the rule to be examined (Parsons, Osherson, & Martinez 1999); and c) ACC which gives an error signal consisting of either the correct answer or an indication that the response was correct or not (Holroyd & Coles 2002). The model itself is of the right inferior frontal cortex where VMPFC and HRR information is combined to select and apply the appropriate transformation, to solve the Wason task (Parsons & Osherson 2001). It is during the application of the transformation that learning also occurs. This biologically plausible network accounts for the difference in the typical correct versus incorrect responses to content-dependent and content-independent versions of the Wason task respectively. In addition, it explains why those trained in logic do better on the content-independent tasks (Rinella, Bringsjord, & Yang 2001).

Model derivations and results

Here we demonstrate how the main computations can be performed in a spiking network using NEF and present some results.

HHRs in spiking networks

Following Plate (), structure is encoded in a representation, using circular convolution ($\otimes$), which implements a kind of vector binding. In order to decode the structure circular correlation ($\oplus$) is used. These operations are defined as:

$$\begin{array}{rclcrcl} \mathbf{C} & = & \mathbf{A}\otimes\mathbf{B} & \quad and \quad & \mathbf{B} & \approx & \mathbf{A}\oplus\mathbf{C} \\ c_j & = & \sum_{k=0}^{n-1} a_k b_{j-k} & & b_j & = & \sum_{k=0}^{n-1} a_k c_{j+k}. \end{array}$$

where subscripts are modulo n. It is often simpler to use the 'involution' or approximate inverse ($'$) in defining transformations of HRRs, where

$$\begin{array}{rcl} \mathbf{A}' & = & (a_o, a_{n-1}, a_{n-2}, \ldots, a_1) \\ \mathbf{I} & \approx & \mathbf{A}'\otimes\mathbf{A}. \end{array}$$

[*] I would like to thank Ray Singh for his preliminary work on the high-level learning algorithm and John Conklin for his invaluable assitance in implementing the circular convolution operation with spiking neurons.

As a result, the following identity holds $\mathbf{A} \oplus \mathbf{B} = \mathbf{A}' \otimes \mathbf{B}$, which is useful since convolution is associative and commutative but correlation is neither.

To implement convolution in a spiking network using the NEF, we first define the encoding and decoding for a vector $\mathbf{x}$:

Encoding

$$a_i(t) = \sum_n \delta(t - t_{in}) = G_i \left[\alpha_i \left\langle \mathbf{x} \cdot \tilde{\phi}_i \right\rangle_m + J_i^{bias}\right]$$

Decoding

$$\hat{\mathbf{x}} = \sum_{i,n} h_i(t - t_n)\phi_i^{\mathbf{x}}$$

where $\delta_i(\cdot)$ are the spikes at times $t_n$for neuron a_i, generated by the nonlinearity G_i(here, a LIF neuron). The α_i is a gain, $\tilde{\phi}_i$ is the preferred direction vector of the neuron in stimulus space, and J_i^{bias} is a bias current that accounts for background activity. For the decoding, $h_i(t)$ are the linear decoding filters which, for reasons of biological plausibility, we take to be the (unweighted) post-synaptic currents (PSCs) in the subsequent neuron. The decoding vectors, $\phi_i^{\mathbf{x}}$, are found by a least-squares method.

Assuming this kind of vector representation in four populations, a, b, c, and d we can now implement the convolution function by using the convolution theorem. First, the two vectors to be convolved are projected through the FFT matrix into a middle layer:

$$\begin{aligned}
&c_k([\mathbf{A}_{FFT}, \mathbf{B}_{FFT}]) \\
&= G_k \left[\alpha_k \tilde{\phi}_k([\mathbf{A}_{FFT}, \mathbf{B}_{FFT}]) + J_k^{bias}\right] \\
&= G_k \left[\textstyle\sum_i \omega_{ik} a_i + \sum_j \omega_{jk} b_j + J_k^{bias}\right]
\end{aligned}$$

where $\omega_{ik} = \alpha_k \tilde{\phi}_{k_1 \dots k_N} \mathbf{W}_{FFT} \phi_i^{\mathbf{A}}$. Then the element-wise product is extracted and the IFFT is performed, giving

$$\begin{aligned}
&d_l(\mathbf{A} \otimes \mathbf{B}) \\
&= G_l \left[\alpha_l \tilde{\phi}_l(\mathbf{W}_{IFFT}(\mathbf{W}_{FFT}\mathbf{A}.\mathbf{W}_{FFT}\mathbf{B})) + J_k^{bias}\right] \\
&= G_l \left[\alpha_l \left(\tilde{\phi}_l \mathbf{W}_{IFFT} \textstyle\sum_k c_k \phi_k^{A.B}\right) + J_k^{bias}\right] \\
&= G_l \left[\textstyle\sum_k \omega_{lk} c_k + J_l^{bias}\right]
\end{aligned}$$

where $\omega_{lk} = \alpha_l \tilde{\phi}_l \mathbf{W}_{IFFT} \phi_k^{A.B}$. This derivation demonstrates how spiking networks, without nonlinear dendritic interactions can compute complex, nonlinear functions like convolution. The implementation of convolution in a spiking network is shown to be successful in figure 1.

Learning HRR transformations

In order to explain the results of the Wason task, we need to transform HRRs encoding the rule being examined appropriately. For the example rule "If there is a vowel on one side of the card, there is an even number on the other side", we can write: $\mathbf{R} = \mathbf{ante} \otimes \mathbf{vowel} + \mathbf{rel} \otimes \mathbf{impl} + \mathbf{cons} \otimes \mathbf{even}$. The typical answer for this 'abstract' Wason task is for subjects to choose to turn over the cards with vowels and with even numbers. A transformation of the rule that provides these results would be $\mathbf{T}_1 = \mathbf{ante}' + \mathbf{impl}' \otimes \mathbf{rel}' + \mathbf{conseq}'$, since $\mathbf{T}_1 \otimes \mathbf{R} \approx \mathbf{vowel} + \mathbf{even}$. However, in the more familiar 'permissive' task, signalled in the model by differing activity from the VMPFC, subjects tend to choose the correct answer (vowel and not even).

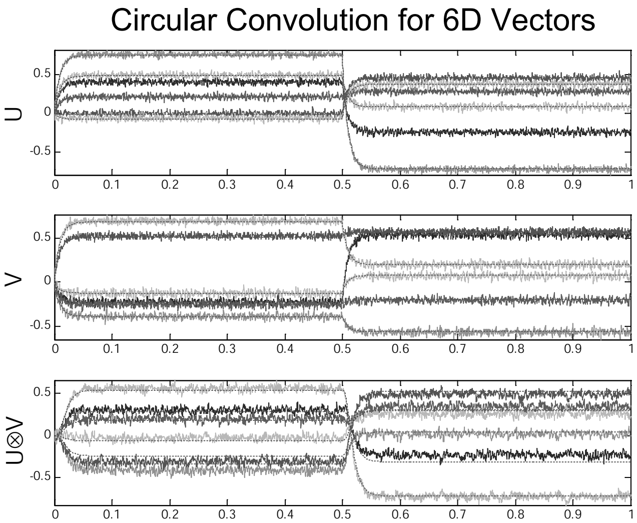

Figure 1: Convolution of two 6 dimensional vectors in a spiking network.

To model learning of these different transformations in different contexts, we extend the work of Neumann (2001). She noted that to find some unknown transformation $\mathbf{T}$ between two vectors $\mathbf{A}$ and $\mathbf{B}$, we can solve $\mathbf{T} = circ(\sum_i^m \mathbf{B}_i \oplus \mathbf{B}_i)^{-1}(\sum_i^m \mathbf{B}_i \oplus \mathbf{A}_i)$, where $circ(\cdot)$ is the circulant matrix and m is the number of examples. However, noting $\mathbf{B}_i \oplus \mathbf{B}_i \approx 1$ this can be simplified to $\mathbf{T} = \frac{1}{m}\sum_i^m \mathbf{B}_i \oplus \mathbf{A}_i$. This can be implemented using the standard delta rule $\mathbf{T}_{i+1} = \mathbf{T}_i - w_i(\mathbf{T}_i - \mathbf{A}_i \oplus \mathbf{B}_i)$, where w_i is an adaptive learning rate inversely proportional to i. This rule needs to be written in terms of individual neural firing. The somewhat involved derivation results in the rule:

$$\begin{aligned}
&\Delta w_{jl} = \kappa \tfrac{\delta E}{\delta \omega_{jl}} \\
&= \tfrac{\kappa}{\alpha_j \|\tilde{\phi}_j\|} \left[\textstyle\sum_k \omega_{jk} z_k - \sum_{j'} \omega_{j'j} y_j\right] (y_j > 0) x_l
\end{aligned}$$

As demonstrated in figure 2, this rule leads to succesful learning, and allows for the switching of learned transformations.

Results

The entire network consists of nine interconnected populations for a total of approximately twenty thousand neurons. This network is able to reproduce the typical results from the Wason task under both the abstract and permissive contexts, as shown in figure 3.

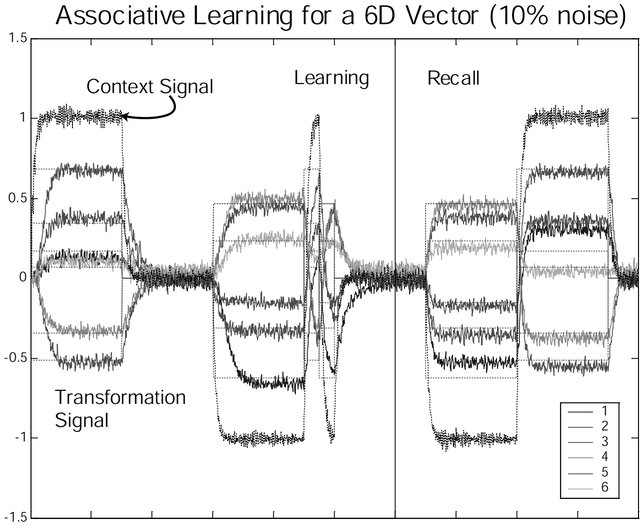

Figure 2: Learning of two different 6 dimensional vectors under two different contexts.

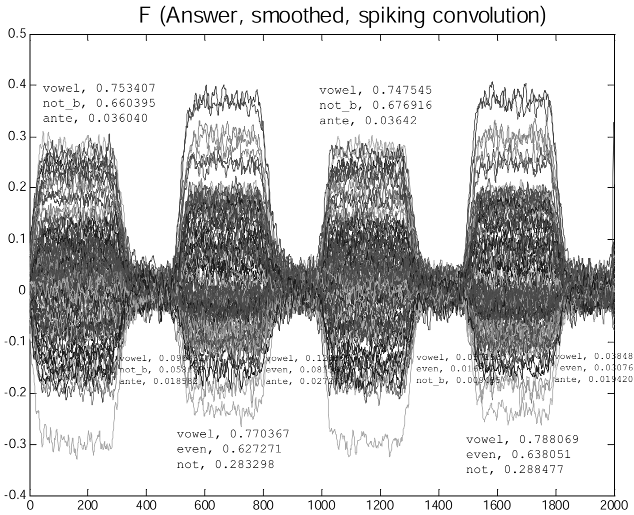

Figure 3: The results from the entire network. The answers are type written above the decoded neural output. As expected, during the first phase (after a brief learning period), in the permissive context, the correct responses are given. In the second phase (a 'no task' period) no answer is given, and in the third phase, in the abstract context, the wrong responses are given. This pattern is repeated twice, to demonstrate that the learning is not destructive.

References

Adolphs, R.; Bechara, A.; Tranel, D.; Damasio, H.; and Damasio, A. 1995. Neuropsychological approaches to reasoning and decision-making. In Damasio, A.; Damasio, H.; and Christen, Y., eds., *Neurobiology of Decision-Making*. New York: Springer Verlag.

Cheng, P. W., and Holyoak, K. J. 1985. Pragmatic reasoning schemas. *Cognitive Psychology* 17:391–416.

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

Eliasmith, C., and Thagard, P. 2001. Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science* 25:245–286.

Fodor, J., and Pylyshyn, Z. 1988. Connectionism and cognitive science: A critical analysis. *Behavioral and Brain Sciences* 28:3–71.

Holroyd, C., and Coles, M. 2002. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review* 109:679–709.

Jackendoff, R. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Kanerva, P. 1994. The spatter code for encoding concepts at many levels. *Proceedings of International Conference on Artificial Neural Networks* 46:226–229.

Neumann, J. 2001. *Holistic Processing of Hierarchical Structures in Connectionist Networks*. PhD dissertation, University of Edinburgh, Department of Computer Science.

Parsons, L., and Osherson, D. 2001. New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex* 11:954–965.

Parsons, L.; Osherson, D.; and Martinez, M. 1999. Distinct neural mechanisms for propositional logic and probabilistic reasoning. *Proceedings of the Psychonomic Society Meeting* 61–62.

Plate, A. Holographic reduced representations: Convolution algebra for compositional distributed representations. In Mylopoulos, J., and Reiter, R., eds., *Proceedings of the 12th International Joint Conference on Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann.

Rinella, K.; Bringsjord, S.; and Yang, Y. 2001. Efficacious logic instruction: People are not irremediably poor deductive reasoners. In Moore, J., and Stenning, K., eds., *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, 851–856. Mahwah, NJ: Lawrence Erlbaum Associates.

Smolensky, P. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence* 46:159–217.

Wason, P. C. 1966. Reasoning. In Foss, B. M., ed., *New horizons in psychology*. Harmondsworth: Penguin.

Wharton, C., and Grafman, J. 1998. Deductive reasoning and the brain. *Trends in Cognitive Sciences* 2:54–59.