An Infomax Algorithm Can Perform Both Familiarity Discrimination and Feature Extraction in a Single Network

Andrew Lulham
MRC Centre for Synaptic Plasticity, Department of Anatomy, and Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K.

Rafal Bogacz
R.Bogacz@bristol.ac.uk
Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K.

Simon Vogt
vogt@isip.uni-luebeck.de
Department of Computer Science, University of Bristol, Bristol BS8 1UB, U.K., and Institute for Signal Processing, University of Lübeck, D-23562 Lübeck, Germany

Malcolm W. Brown
M.W.Brown@bristol.ac.uk
MRC Centre for Synaptic Plasticity, Department of Anatomy, University of Bristol, Bristol BS8 1UB, U.K.

Psychological experiments have shown that the capacity of the brain for discriminating visual stimuli as novel or familiar is almost limitless. Neurobiological studies have established that the perirhinal cortex is critically involved in both familiarity discrimination and feature extraction. However, opinion is divided as to whether these two processes are performed by the same neurons. Previously proposed models have been unable to simultaneously extract features and discriminate familiarity for large numbers of stimuli. We show that a well-known model of visual feature extraction, Infomax, can simultaneously perform familiarity discrimination and feature extraction efficiently. This model has a significantly larger capacity than previously proposed models combining these two processes, particularly when correlation exists between inputs, as is the case in the perirhinal cortex. Furthermore, we show that once the model fully extracts features, its ability to perform familiarity discrimination increases markedly.

1 Introduction

Familiarity discrimination is concerned with determining whether a stimulus is novel or has been previously encountered. Investigations into the capacity of the human brain for familiarity discrimination have shown that even after single-trial presentations of 10,000 pictures, participants were able to discriminate their familiarity with an average accuracy of 83% (Standing, 1973).

Results obtained using various experimental techniques (including studies of amnesic patients, lesion studies, fMRI, single neuron recording, and gene expression) have established that familiarity discrimination is critically dependent on a part of the medial temporal lobe called the *perirhinal cortex* (for reviews, see, Brown & Aggleton, 2001; Brown & Xiang, 1998; Eichenbaum, Yonelinas, & Ranganath, 2007; Murray & Bussey, 1999). Furthermore, these data indicate that a fraction of perirhinal neurons are capable of discriminating novel and familiar visual stimuli by a difference in firing rate—namely, the neurons have a reduced firing rate for stimuli that have previously been seen (Brown, Wilson, & Riches, 1987; Brown & Xiang, 1998; Fahy, Riches, & Brown, 1993; Li, Miller, & Desimone, 1993; Miller, Li, & Desimone, 1993; Riches, Wilson, & Brown, 1991; Sobotka & Ringo, 1993; Xiang & Brown, 1998). Such neurons are called novelty neurons, and they have been found in the perirhinal and neighboring cortices (Brown & Xiang, 1998).

As well as familiarity discrimination, the perirhinal cortex is also involved in visual processing, and experiments suggest that neurons in this area represent conjunctions of features of visual stimuli (Bussey, Saksida, & Murray, 2005; Murray & Bussey, 1999). Recordings from neurons in the inferotemporal cortex, of which the perirhinal cortex is part, have revealed that they alter their patterns of responsiveness to sets of familiar stimuli after the addition of novel stimuli (Kobatake, Wang, & Tanaka, 1998; Rolls, Baylis, Hasselmo, & Nalwa, 1989). This finding shows that representations of stimuli in the perirhinal cortex are not constant but change to incorporate new information. This is a strong indicator that feature extraction is taking place. A model of the medial temporal lobe has been proposed that incorporates the perirhinal cortex in this role (Bussey et al., 2005).

Given these findings, models have been proposed that attempt to combine the perceptual and mnemonic roles of the perirhinal cortex (Norman, Newman, & Perotte, 2005; Norman & O'Reilly, 2003; Sohal & Hasselmo, 2000). However, it has been shown that simplified versions of two of these models, both based on Hebbian learning (Norman & O'Reilly, 2003; Sohal & Hasselmo, 2000), have a greatly reduced capacity for familiarity discrimination when the inputs to the networks are correlated (Bogacz & Brown, 2003; see section 3.1 for details of patterns used in these testing these models), a condition shown to be realistic in the perirhinal cortex (Erickson, Jagadeesh, & Desimone, 2000). It has also been shown (Bogacz & Brown, 2003) that these two "combined" models (Norman & O'Reilly, 2003; Sohal & Hasselmo, 2000) fail to extract independent features. The model of Norman et al. (2005), which combines Hebbian and anti-Hebbian learning, can achieve high capacity for familiarity discrimination, but we are not aware of any published analysis of its ability to extract features.

Another model that has high capacity for familiarity discrimination (proportional to the number of synapses in the network) is based on anti-Hebbian learning (Bogacz & Brown, 2003). Additionally, this model reproduces the reduced neuronal response to familiar stimuli seen in recordings from perirhinal neurons. However, this model is specialized in familiarity discrimination, positing that a separate network of perirhinal neurons performs feature extraction. Until now, it has not been known if it is possible to efficiently perform familiarity discrimination and feature extraction within a single neural network.

Here we show that a well-known model of visual feature extraction, Infomax (Bell & Sejnowski, 1995), can simultaneously perform familiarity discrimination and feature extraction efficiently. This and similar algorithms have been applied to modeling visual feature extraction in the primary visual cortex and have been shown to result in neurons with receptive fields similar to those observed experimentally (Bell & Sejnowski, 1997; Bogacz, Brown, & Giraud-Carrier, 2001a; Olshausen & Field, 1996, 1997). A similar algorithm has been also used to model feature extraction in later stages of ventral visual stream (Waydo & Koch, 2008). The Infomax model includes an anti-Hebbian term in its learning rule, and its capacity for familiarity discrimination scales with network size in a similar way to the anti-Hebbian model. Since Infomax is a relatively abstract model, we do not propose that it describes the details of information processing in neural circuits of the perirhinal cortex, but rather that it suggests general computational principles that could be employed by this cortex to achieve high-efficiency familiarity discrimination combined with feature extraction.

In section 2 we give details of the Infomax model. Sections 3 and 4 then describe simulations performed to test the capacity of the Infomax model, first for familiarity discrimination alone and then for familiarity discrimination combined with feature extraction. In section 5, we discuss possible future directions for investigating how the Infomax model could be implemented in the perirhinal cortex and the relationship of the model to experimental data.

Simulations establishing model capacity for familiarity discrimination and feature extraction can be replicated using the familiarity discrimination toolbox, which can be downloaded from http://www.cs.bris.ac.uk/Research/MachineLearning/FamTool/.

2 The Infomax Model

The Infomax model is implemented in a fully connected network with N neurons in each layer, as shown in Figure 1. The weights w_{ij} of the connections between input j and novelty neuron i are initialized by randomly generated numbers from a uniform distribution between -0.5 and 0.5 and then normalized such that for each novelty neuron i, the standard deviation of the associated weights is 1 and the mean is 0. We assume that all stimuli

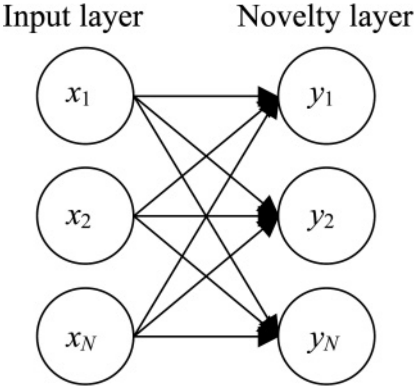

Figure 1: Architecture of the Infomax model. Circles denote neurons, and arrows denote connections. The network is fully connected, with a layer containing novelty neurons receiving feedforward projections from a layer of input neurons.

are represented as patterns of activity of the input neurons (i.e., vectors of length N). During the learning phase, the activation values x_j of input neurons are set to the pattern being learned, and the synaptic inputs of the novelty neurons are computed from

$$h_i = \sum_{j=1}^{N} w_{ij} x_j. \tag{2.1}$$

The activation values y_i of the novelty neurons are then computed as

$$y_i = \tanh(h_i). \tag{2.2}$$

The weights w_{ij} determine how inputs x_j are encoded in the activity of novelty neurons y_i. The Infomax algorithm finds the values of w_{ij} for which y_i maximizes the information about x_j (hence the name of the algorithm). The information is maximized when activities of different novelty neurons are independent. If the activity of two neurons is correlated, they carry less information (in the extreme case of two neurons having fully correlated, i.e., identical, activity, they carry the same amount of information as a single neuron). Hence, Infomax can be used to extract independent features.

In the original formulation of the Infomax algorithm (Bell & Sejnowski, 1995), after presentation of each input pattern x_j, the weights of the connections between neurons are modified according to the following rule so as to optimally improve the information carried by y_i about x_j (i.e., modification is proportional to the gradient of the information over the weights):

$$\Delta w_{i,j} = \frac{\eta}{N}\left((w^T)^{-1}_{i,j} - 2y_i x_j\right). \tag{2.3}$$

Here η represents the learning rate. To understand why the Infomax model can be used for familiarity discrimination, let us consider the second term on the right-hand side, $-2y_i x_i$. It is an anti-Hebbian term, so called because it works in the opposite way to the Hebbian learning rule (Hebb, 1949): the minus sign means that the weights between coactive neurons are weakened rather than strengthened. This tends to cause a response closer to zero for familiar stimuli, because when a stimulus is repeated, less input is received through the weakened weights. This property is important because it allows the model to discriminate familiarity on the basis of neuronal responses (as described toward the end of this section). Such an anti-Hebbian term has been used in previous models of familiarity discrimination (Bogacz & Brown, 2003; Brown & Xiang, 1998; Kohonen, 1989). Anti-Hebbian weakening of synaptic weights between coactive neurons is biologically plausible because homosynaptic LTD can be demonstrated in the perirhinal cortex (Brown & Bashir, 2002).

The first term of equation 2.3, $(w^T)^{-1}_{i,j}$, involves the calculation of an inverse matrix, which is computationally expensive; hence, in our simulations, the weights are updated using the following extended learning rule (Lee, Girolami, & Sejnowski, 1999):

$$\Delta w_{i,j} = \frac{\eta}{N}\left(w_{i,j} - (y_i + h_i)\sum_{k=1}^{N} h_k w_{k,j}\right). \tag{2.4}$$

This rule uses the natural gradient weight update (Amari, Cichocki, & Yang, 1996) and was specially designed for patterns generated by combining features so that each feature occurred in only a small fraction of patterns. The patterns we used in our simulations have this property (see section 3.1), and hence in this letter, we present the results obtained with this extended rule.

Since in the Infomax model, the responses of novelty neurons tend to be closer to zero for repeated stimuli, a decision on the familiarity of a presented stimulus may be reached by measuring the overall response of novelty neurons. Hence, familiarity discrimination is simulated in the model in the following way. The activities of input neurons are set to the pattern being discriminated. The synaptic inputs are computed from equation 2.1, and the total response of the network on presentation of a stimulus x is computed as

$$d(x) = \sum_{i=1}^{N} |h_i|, \tag{2.5}$$

where $|\ |$ denotes the absolute value. We refer to d as the decision function. If the decision function is above a certain threshold (determined as described in section 3.1), the pattern is classified as novel. Otherwise, it is classified as familiar.

3 Capacity of the Infomax Model for Familiarity Discrimination

3.1 Simulation Method. In order to compare models, Bogacz and Brown (2003) developed benchmark tests for measuring the capacity of models for familiarity discrimination. These tests have been used to compare and contrast a number of previous models of familiarity discrimination, so for comparison, the same tests were used here.

The capacity is defined as the number of stored patterns P for which the familiarity discrimination error rate is equal to 1%. We search for this capacity by evaluating error rate for different values of P in the following way. For a given value of P, a set of $2P$ stimuli is generated. Each stimulus is a pattern of activity of N input neurons (a vector of length N). This set of stimuli is split, and each pattern from the first P of the stimuli is presented once to the model. These P training stimuli should then be "familiar" to the network, so that if we present them again, the values of the decision function for each of them should be lower than for the novel, untrained stimuli.

The network is then tested by presenting all $2P$ of the stimuli—both trained and untrained—to the network and evaluating the decision function for each. A threshold value that best separates the decision function values of the two groups is then found numerically. The number of incorrect decisions divided by the total number of stimuli gives the error rate.

It is necessary to establish the resistance of models to correlation between input neurons, since it has been shown that this type of correlation exists in the perirhinal cortex (Erickson et al., 2000). We tested the Infomax model on two types of patterns with such correlations. The first type of patterns (type 1) were those previously used by Bogacz and Brown (2003) to test the capacity of other models' familiarity discrimination. These patterns are generated by introducing a specified amount of correlation between each pair of input neurons. This is done by first generating an initial template pattern, x^{temp}, by randomly assigning each of its bits x_j^{temp} to 1 or –1. Subsequent patterns are then generated such that the bit at position j is equal to x_j^{temp} with probability $^1/_2 + {}^1/_2b$, or is equal to $-x_j^{temp}$ with probability $^1/_2 - {}^1/_2b$. The parameter b controls the similarity between patterns and the template, and it determines the amount of correlation between the activities of input neurons, as b^2 is equal to the absolute value of correlation $|r_{ij}|$ between x_i and x_j across the patterns (in particular, $r_{ij} = b^2$ if $x_i^{temp} = x_j^{temp}$, and $r_{ij} = -b^2$ if $x_i^{temp} \neq x_j^{temp}$). Simulations were performed for values of b between 0 and 0.9 at intervals of 0.1. In order to ensure each bit position is on average equally active, it is necessary to invert half of them after generating patterns.

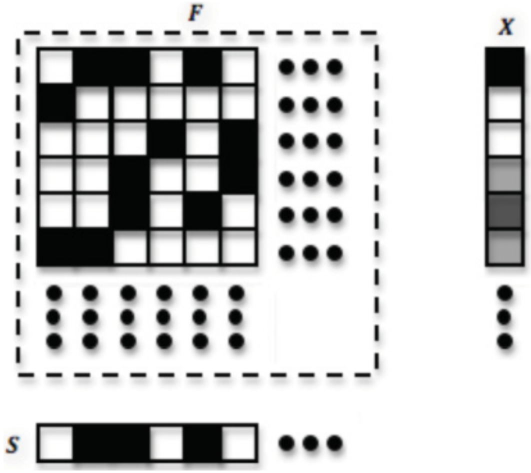

Figure 2: An example showing how a feature-based pattern is generated. The matrix inside the dashed square is a sample feature matrix. Each column in this matrix is a feature, and empty and filled squares denote bits equal to 0 and 1, respectively. For simplicity, only the first bits of a few first features are shown. The vector at the bottom of the figure indicates which features are used to create the pattern (it is the source vector). In particular, its second, third, and fifth bits are equal to 1, which implies that the pattern is created by adding the second, third, and fifth features. The resulting pattern is shown on the right, and different shades of gray denote different values of x_i (0 = white, light gray = 1, dark gray = 2, black = 3). Thus, for example, $x_1 = 3$ because the first bit is equal to 1 in all three features from which the pattern is composed. Analogously, $x_2 = x_3 = 0$ because the second and third bits are equal to 0 in the three features, while $x_4 = 1$ because the fourth bit is equal to 1 only in the third feature.

The second type of patterns (type 2) is generated by combining multiple features and was previously used by Bogacz and Brown (2003) to assess the ability of models to perform feature extraction. The patterns are generated in the following way. First, a set of M sparse features, f, described by binary vectors of length N, is generated independent of one another. These features are used as building blocks to construct the patterns, x. Each pattern is formed by mixing a fixed number of randomly chosen features from the feature set. Let $f_{j,i}$ be the jth bit of feature i. Let $s_{i,\mu}$ indicate if feature i is present in pattern μ. The patterns are generated according to

$$x_{j,\mu} = \sum_{i=1}^{M} f_{j,i} s_{i,\mu}. \tag{3.1}$$

Figure 2 illustrates how a sample pattern is formed. We refer to $f_{j,i}$ and $s_{i,\mu}$ as features and sources, but sometimes they are referred to in the literature as factors and loadings, respectively. We fix the number of ones in each source vector s_μ to $N/10$, meaning that each pattern generated contains $N/10$ features. The number of bits switched on in each feature is set to 30 (the model is, however, robust to changes in this parameter). Input neurons are also constrained to be active for an equal number of features (i.e., $\sum_{i=1}^{M} f_{j,i} = const$).

3.2 Results. Figure 3a shows the simulated capacity of the Infomax model for patterns similar to a template (i.e., type 1) with different levels of similarity b and numbers of neurons per layer, N. For comparison, Figure 3b shows the capacity of the anti-Hebbian model (Bogacz & Brown, 2003). Although there are quantitative differences between the two models, it can be seen from Figures 3a and 3b that the Infomax and anti-Hebbian models show qualitatively similar dependence of capacity on b and N. This is to be expected due to the similarity in the learning rules for these two models.

We investigated how the capacity of the Infomax model scales with network size, for a realistic level of correlation between input neurons. Erickson et al. (2000) found the mean correlation between distant pairs of neurons in the perirhinal cortex to be 0.04. As Bogacz and Brown (2003) showed, two previously published combined models were unable to perform familiarity discrimination efficiently at this level of correlation. Figure 3c shows the capacity of the Infomax model if we fix the similarity at $b = 0.2$ (equivalent to a correlation of 0.04) and vary the network size. The capacity of the model is shown together with the closest quadratic fit to the results of simulations, given by the following equation:

$$P = 0.0046N^2 + 0.66N - 0.74. \tag{3.2}$$

Bogacz and Brown (2003) have shown that the capacity of the anti-Hebbian model scales quadratically with network size ($P = \mathrm{O}(N^2\ )$). Here we see the same is true of the Infomax model.

Figure 3d shows the capacity of the Infomax model for the feature-based patterns (type 2) as a function of the network size, together with the closest linear and quadratic fits. The quadratic fit provides a better match for the data ($p < 10^{-5}$, F-test for nested models) and is given by the following equation:

$$P = 0.0034N^2 + 1.2N - 96. \tag{3.3}$$

In summary, Figures 3a to 3c show that the capacity of the Infomax model scales similarly to that of the anti-Hebbian model. Furthermore, the Infomax model is also resistant to correlated firing between input neurons, in particular at realistic levels.

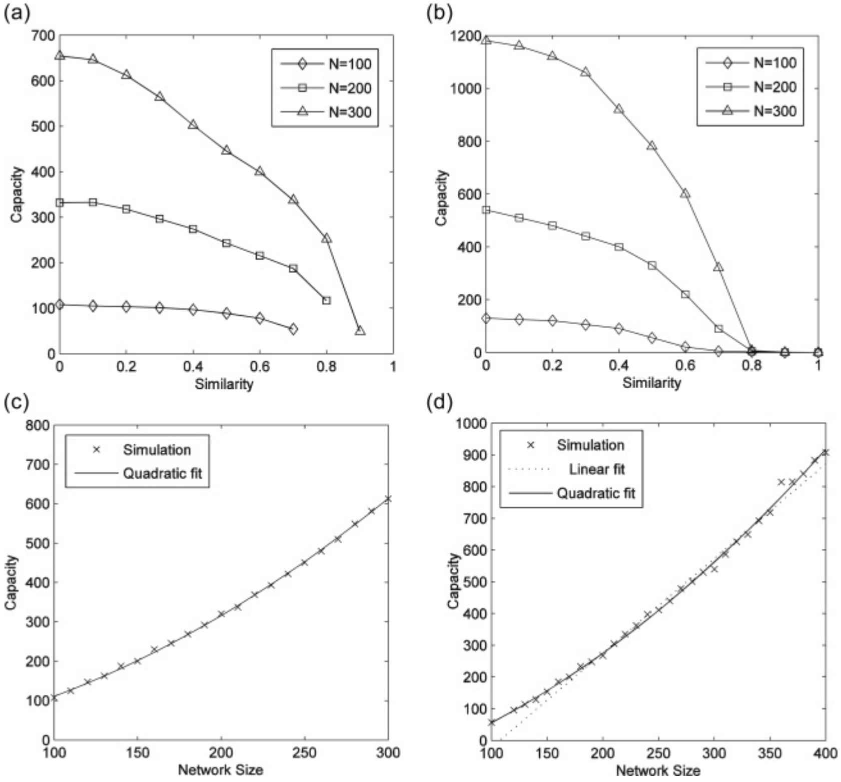

Figure 3: Capacity for familiarity discrimination. Solid lines show the number of patterns the network could successfully discriminate with an error rate of ≤ 1%, as found by simulations. (a) The capacity of the Infomax model for three different network sizes (*N*) for different values of the similarity of the input patterns to the template. Error bars are not shown (the maximum standard error was 4.2). (b) The capacity of the anti-Hebbian model. Data taken from Figure 5d of paper by Bogacz and Brown (2003). (c) The capacity of the Infomax model as a function of network size for patterns with a similarity to template equal to 0.2. (d) The capacity of the Infomax model as a function of network size for feature-based patterns.

4 Simultaneous Familiarity Discrimination and Feature Extraction

We know that Infomax in its original, iterative form is a feature extraction algorithm. However, it remains to be tested whether the algorithm can perform both familiarity discrimination and feature extraction simultaneously, or even whether it can still perform feature extraction when used in a single-trial fashion. Both of these questions are resolved in this section.

4.1 Method of Simulation. In order to measure how many features a model can extract, we again use the feature-based patterns (type 2, described in section 3.1). The generated patterns are presented to the network in training blocks of 5000. After each training block, the familiarity discrimination of the network is tested using t novel patterns (generated from the same feature matrix) and the t most recent familiar patterns, where $t \in \{500, 5000\}$. This test is carried out as described in section 3.1 and produces an error rate value (or, equivalently, an accuracy value). Unlike in the previous familiarity discrimination simulations, we do not search for a capacity giving a 1% error rate.

We then test the number of features extracted. In order to understand the method used, it is important to note that equation 3.1 can be expressed in matrix form as

$$X = F \times S. \tag{4.1}$$

F, S, and X are matrices composed of columns of features f_j, sources s_μ, and patterns x_μ, respectively. The goal of feature extraction is then to modify network weights so that on presentation of any pattern x_μ, the synaptic inputs of novelty neurons reflect the respective source s_μ. In matrix form, this can be expressed as

$$W \times X = P_\pi \times S, \tag{4.2}$$

where W is the weight matrix. Since in feature extraction it does not matter which novelty neuron corresponds to which source s_i (e.g., h_i does not need to be equal to s_i, but there should exist a k such that $h_k = s_i$), the source matrix S in equation 4.2 is multiplied by a permutation matrix P_π (in which each row and each column contains a single 1 and is otherwise filled with zeros, so that $P_\pi \times S$ is equivalent to the matrix S but with permuted rows). Substituting equation 4.1 into equation 4.2, canceling S, and post-multiplying by the inverse of F gives us

$$W = P_\pi \times F^{-1}. \tag{4.3}$$

Equation 4.3 is the basis for how feature extraction is assessed. If features have been fully extracted, the weight matrix should be a row permutation of the inverse of the feature matrix.

We use the following technique to measure feature extraction. For each row in the inverse feature matrix, we find the most correlated row in the weight matrix (using the absolute values of Pearson correlation coefficients) and then take the mean of these absolute values of correlation. For perfect feature extraction, this mean value will equal 1. For the initial (randomly generated) weights, the expected value of the feature extraction measure depends on network size and is equal to 0.21 for $N = 100$, 0.16 for $N = 200$, and 0.12 for $N = 300$.

We continue presenting blocks of training patterns until 500 blocks (2.5 $\times 10^6$ patterns) have been presented. The tests were repeated for a range of learning rates and network sizes.

4.2 Results. Figure 4 shows the results of the simultaneous familiarity discrimination and feature extraction simulations for two learning rates and three network sizes. For all parameters shown, the measure of feature extraction always asymptotes. The number of presented patterns required for convergence is larger for larger networks, since the number of features increases with network size so the problem becomes more difficult. For the higher learning rate (see Figure 4, right panels), the convergence of feature extraction is much faster than for the lower learning rate (see Figure 4, left panels), but the value to which the feature extraction converges is always lower, indicating convergence to a poorer quality set of features.

Across learning rates and test sizes, the accuracy for familiarity discrimination increases with increasing network size. It is also higher for 500 test patterns than for 5000, due to forgetting; since in the model, the information about all familiar patterns is stored in the same set of synapses, the information about older patterns may be overwritten by newer ones. For the higher learning rate (see Figure 4, right panels), forgetting happens more rapidly than for the lower learning rate (see Figure 4, left panels), as shown by the larger loss in accuracy for the larger compared to the smaller test size.

For the lower learning rate (see Figure 4, left panels), the familiarity discrimination accuracy increases once features have been extracted. This indicates that accurately extracting underlying features can in some cases boost accuracy for familiarity discrimination.

Importantly, the model simultaneously discriminates familiarity with high accuracy and extracts features. In particular, for 300 neurons and a learning rate of 0.05 (bottom-left panel), not only are features fully extracted, but the familiarity discrimination accuracy for 500 test patterns, is 100%, and for 5000 test patterns, it is in excess of the 83% accuracy found experimentally for humans presented with 5000 stimuli (Standing, 1973).

5 Discussion

We have shown that the Infomax model is able to perform familiarity discrimination efficiently and is resistant to correlated firing between input neurons (in particular, to levels of correlation shown to be prevalent in the perirhinal cortex). Elsewhere it has also been shown that the Infomax model is able to reproduce Standing's presented and retained familiarity discrimination power law (Androulidakis, Lulham, Bogacz, & Brown, 2008). Moreover, the model is able to simultaneously perform familiarity discrimination and feature extraction.

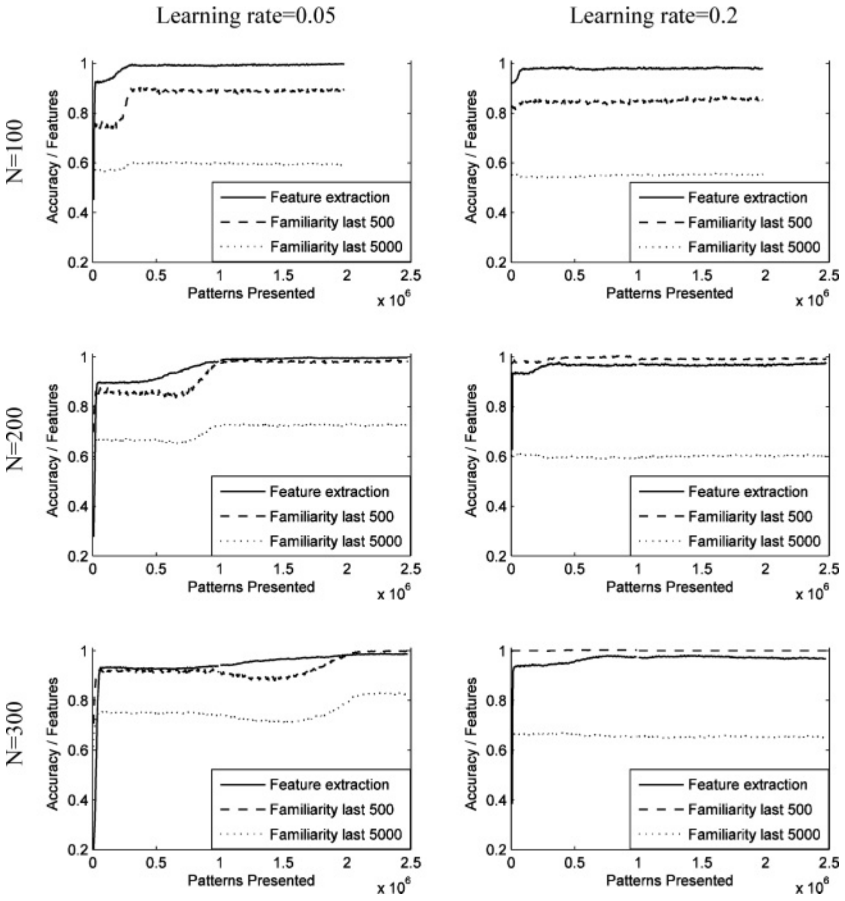

Figure 4: Simultaneous familiarity discrimination and feature extraction. For each panel, accuracy is shown by dashed and dotted lines and percentage feature correlation by a solid line. The top row of panels shows results for a network of 100 neurons, the middle row shows results for 200 neurons, and the bottom row for 300 neurons. Panels in the left column use a lower learning rate of 0.05, whereas panels in the right column use a higher learning rate of 0.2. For all panels, the dotted line shows the accuracy when testing on the 5000 most recently presented patterns, whereas the dashed line shows the accuracy when testing on the 500 most recently presented patterns. Patterns were generated with 30 bits per feature and $N/10$ features per pattern.

5.1 How Is Simultaneous Familiarity Discrimination and Feature Extraction Possible? Simultaneous familiarity discrimination and feature extraction is intuitively a particularly challenging task, since the requirements for the two goals are seemingly conflicting. Familiarity discrimination requires fast one-shot learning, whereas feature extraction requires gradual learning over multiple stimulus presentations. Nevertheless, it appears that with Infomax, learning rates exist for which the two processes are mutually compatible.

If biological feature extraction networks are to be adaptive to changes in an environment, they need to extract features throughout the animal's life, and hence their learning rates should never decay to zero. The learning rate η does not decay in our model; hence, the weights in the network never converge, even if our measures of feature extraction are very close to 1 (as in the bottom-left panel of Figure 4). Instead, the weights fluctuate around the values required for perfect feature extraction (see equation 4.3) because they are modified after each pattern presentation. These small departures from equation 4.3 encode information sufficient to discriminate which patterns have been presented to the network.

5.2 Influence of Feature Extraction on Familiarity Discrimination. Our simulations also demonstrate that once the features are extracted, the accuracy of familiarity discrimination may improve. We now discuss a possible reason for this observation. Before the weights converge to the vicinity of the values of equation 4.3, after each pattern presentation they are modified such that they are moved closer to the values of equation 4.3, and they encode information specific to the presented pattern. By contrast, once the weights are in the vicinity of the values of equation 4.3, after each pattern presentation, they are modified such that they encode information specific to the presented pattern. Thus, in the latter case, a larger proportion of weight modifications encodes information specific to individual patterns, which potentially increases the accuracy of familiarity discrimination.

One can observe in Figure 4 that just before the measure of feature extraction asymptotes, the familiarity discrimination accuracy slightly decreases before again increasing toward the value at asymptote. This nonmonotonic behavior of the accuracy is particularly visible in the bottom-left panel of Figure 4, and we now discuss a possible reason for this behavior. The magnitude of weight modification after each stimulus presentation is likely to be higher when the weights are far from the values required for perfect feature extraction (see equation 4.3) than when the weights approach these values (because the gradient of the information maximized by Infomax is likely to be steeper further from the maximum). Thus, as the weights start to approach equation 4.3, less information about the presented stimuli is encoded in the weights than was encoded for the preceding stimuli (for which the weight modifications were larger). Hence, the preceding stimuli interfere with recently presented ones, and the familiarity accuracy for the recently presented stimuli starts to decrease (this aspect of the model has potential parallels in the well-documented effects of proactive interference in reducing the memory performance of subjects). As learning progresses, the memory of these preceding stimuli decays, so they no longer interfere with newly acquired stimuli. Due to this reduced interference and the reduced weight changes associated with feature extraction described in the previous paragraph, the familiarity accuracy increases.

5.3 Relationship to Experimental Data. If a similar algorithm were used for feature extraction throughout the visual stream, then earlier visual areas should also be capable of familiarity discrimination. However, neurons in early visual areas have small receptive fields and are typically activated by simpler visual features; they might therefore be expected to discriminate between novel features rather than complex stimuli. The simpler the type of feature, the less likely it is that a novel one will be encountered and, correspondingly, the role of that part of the network in familiarity discrimination will diminish.

Apart from novelty neurons, neurons with other types of response have also been recorded in the perirhinal cortex (Xiang & Brown, 1998). Recency neurons have a weakened response to stimuli that were recently presented but a strong response to other stimuli (regardless of how familiar the stimuli are to the animal). These pose a new problem when modeling the perirhinal cortex as a whole. Are these neurons part of a separate system, or a subsystem of familiarity discrimination (Bogacz, Brown, & Giraud-Carrier, 2001b)? Whether or not a subset of neurons in the Infomax model behaves as recency neurons (i.e., they have responses closer to zero for recently presented patterns) is currently unclear, but one potential future direction would be to inspect and classify each neuron during the testing phase. If recency neurons were not found, introducing shorter-term plasticity into the model might enable recency neurons to emerge. Alternatively, a separate system with the same learning rule but a higher rate of learning could accurately replicate these recency neurons, but it is unclear what feature extraction would mean in this context.

The Infomax model does not assume any spatial arrangement of the novelty neurons, and hence it cannot account for the data showing an increase in correlation between adjacent neurons for familiar stimuli (Erickson et al., 2000). To address such data, the learning rule would have to be extended so that adjacent neurons represent similar features (as in Cowell, Bussey, & Saksida, 2006), and this would be an interesting direction for future work.

5.4 How Might Infomax Be Implemented in the Perirhinal Cortex? Another important piece of future work that is beyond the scope of this letter would be to investigate how the computational principles of the Infomax model could be implemented by real neurons in the perirhinal cortical network. A number of different feature extraction models could be explored here. For example, the model of Olshausen and Field (1996) performs similar computations to Infomax but is more easily neurally implemented because "the dynamics . . . as well as the learning rule . . . have a local network implementation" (p. 608). Thus, future work could involve investigating whether such more biologically plausible models of feature extraction could also discriminate familiarity efficiently and how the neuronal activity that these models predict in recognition memory tasks relates to neurophysiological data.

In the Infomax model, the synaptic weights of a single neuron can take both positive and negative values and the level of neuronal activity can be negative, which is not biologically plausible. Furthermore, these two properties allow the neurons to encode the presence of a feature in negative activity. When developing a model of feature extraction and familiarity discrimination in the perirhinal cortex, it will be important to constrain it such that the presence of a feature can be encoded only by increased (rather than decreased) activity of neurons representing this feature. In such a model, in which the levels of neuronal activity will be constrained to positive values, the decision function of equation 2.5 will express the total level of neuronal activity (because $|h_i| = h_i$ for $h_i \geq 0$). Thus, such a model would be able to produce reduced levels of neuronal activity for familiar patterns, seen in the perirhinal cortex (see section 1).

6 Conclusion

Work in psychology and neuroscience suggests that there exists a separate familiarity discrimination process in addition to recollection (Brown & Aggleton, 2001; Yonelinas, 2002). Computational models have addressed the question of why the brain would include such an additional familiarity process. Bogacz, Brown, and Giraud-Carrier (2001c) showed that a neural network can perform familiarity discrimination for many more stimuli than could be recollected by an associative memory network of the same size. This suggests that relatively few resources are required to support familiarity discrimination. Here we demonstrate that theoretically, familiarity discrimination can be efficiently performed by the same network that underlies feature extraction. Therefore, the benefits of familiarity discrimination might potentially be achieved with almost no additional resources.

Acknowledgments

This work was supported by an MRC Capacity Building Research Studentship held by Andrew Lulham.

References

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing Systems, 8* (pp. 757–763). Cambridge, MA: MIT Press.

Androulidakis, Z., Lulham, A., Bogacz, R., & Brown, M. W. (2008). Computational models can replicate the capacity of human recognition memory. *Network: Computation in Neural Systems, 19*, 161–182.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research, 37*(23), 3327–3338.

Bogacz, R., & Brown, M. W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus, 13*(4), 494–524.

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001a). Emergence of motion-sensitive neurons' properties by learning sparse code for natural moving images. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 838–844). Cambridge, MA: MIT Press.

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001b). Model of co-operation between recency, familiarity and novelty neurons in the perirhinal cortex. *Neurocomputing, 38*, 1121–1126.

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001c). Model of familiarity discrimination in the perirhinal cortex. *J. Comput. Neurosci., 10*(1), 5–23.

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience, 2*, 51–62.

Brown, M. W., & Bashir, Z. I. (2002). Evidence concerning how neurons of the perirhinal cortex may effect familiarity discrimination. *Philosophical Transactions of the Royal Society of London Series B–Biological Sciences, 357*(1424), 1083–1095.

Brown, M. W., Wilson, F. A. W., & Riches, I. P. (1987). Neuronal evidence that inferomedial temporal cortex is more important than hippocampus in certain processes underlying recognition memory. *Brain Research, 409*(1), 158–162.

Brown, M. W., & Xiang, J. Z. (1998). Recognition memory: Neuronal substrates of the judgement of prior occurrence. *Progress in Neurobiology, 55*(2), 149–189.

Bussey, T. J., Saksida, L. M., & Murray, E. A. (2005). The perceptual-mnemonic/feature conjunction model of perirhinal cortex function. *Quarterly Journal of Experimental Psychology Section B–Comparative and Physiological Psychology, 58*(3–4), 269–282.

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2006). Why does brain damage impair memory? A connectionist model of object recognition memory in perirhinal cortex. *J. Neurosci., 26*(47), 12186–12197.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30*, 123–152.

Erickson, C. A., Jagadeesh, B., & Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkey. *Nature Neuroscience, 3*, 1143–1148.

Fahy, F. L., Riches, I. P., & Brown, M. W. (1993). Neuronal-activity related to visual recognition memory: Long-term-memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal Cortex. *Experimental Brain Research, 96*(3), 457–472.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Hoboken, NJ: Wiley.

Kobatake, E., Wang, G., & Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology, 80*(1), 324–330.

Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.). New York: Springer-Verlag.

Lee, T. W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended Infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation, 11*(2), 417–441.

Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus-familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology, 69*(6), 1918–1929.

Miller, E. K., Li, L., & Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term-memory task. *Journal of Neuroscience, 13*(4), 1460–1478.

Murray, E. A., & Bussey, T. J. (1999). Perceptual-mnemonic functions of the perirhinal cortex. *Trends in Cognitive Sciences, 3*(4), 142–151.

Norman, K. A., Newman, E. L., & Perotte, A. J. (2005). Methods for reducing interference in the complementary learning systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks, 18*(9), 1212–1228.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110*(4), 611–646.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.

Riches, I. P., Wilson, F. A. W., & Brown, M. W. (1991). The effects of visual-stimulation and memory on neurons of the hippocampal-formation and the neighboring parahippocampal gyrus and inferior temporal cortex of the primate. *Journal of Neuroscience, 11*(6), 1763–1779.

Rolls, E. T., Baylis, G. C., Hasselmo, M. E., & Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research, 76*(1), 153–164.

Sobotka, S., & Ringo, J. L. (1993). Investigation of long-term recognition and association memory in unit responses from inferotemporal cortex. *Experimental Brain Research, 96*(1), 28–38.

Sohal, V. S., & Hasselmo, M. E. (2000). A model for experience-dependent changes in the responses of inferotemporal neurons. *Network—Computation in Neural Systems, 11*(3), 169–190.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology, 25*, 207–222.

Waydo, S., & Koch, C. (2008). Unsupervised learning of individuals and categories from images. *Neural Computation, 20*(5), 1165–1178.

Xiang, J. Z., & Brown, M. W. (1998). Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology, 37*(4–5), 657–676.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.

Received February 1, 2010; accepted August 25, 2010.