ARE UNSUPERVISED NEURAL NETWORKS IGNORANT? SIZING THE EFFECT OF ENVIRONMENTAL DISTRIBUTIONS ON UNSUPERVISED LEARNING

Sébastien Hélie[1], Sylvain Chartier[2, 3], & Robert Proulx[1]

[1]Laboratoire d'Étude en Intelligences Naturelle et Artificielle, Université du Québec À Montréal

[2]Université du Québec en Outaouais

[3]Centre de Recherche de l'Institut Philippe Pinel de Montréal

Running head: Are unsupervised neural networks ignorant?

For correspondence:
Sébastien Hélie
Laboratoire d'Étude en Intelligences Naturelle et Artificielle
Département de Psychologie
C. P. 8888, Succ. Centre-ville
Montréal (Québec) H3C 3P8 CANADA

Phone: (514) 987-3257,
E-mail: Helie.Sebastien@courrier.uqam.ca

Keywords: Artificial neural networks, connectionist networks, autoassociative memories, competitive networks, density estimation.

Abstract

Learning environmental biases is a rational behavior: by using prior odds, Bayesian networks rapidly became a benchmark in machine learning. Moreover, a growing body of evidence now suggests that humans are using base rate information. Unsupervised connectionist networks are used in computer science for machine learning and in psychology to model human cognition, but it is unclear whether they are sensitive to prior odds. In this paper, we show that hard competitive learners are unable to use environmental biases while recurrent autoassociative memories use frequency of exemplars and categories independently. Hence, it is concluded that recurrent autoassociative memories are more useful than hard competitive networks to model human cognition and have a higher potential in machine learning.

ARE UNSUPERVISED NEURAL NETWORKS IGNORANT? SIZING THE EFFECT OF ENVIRONMENTAL DISTRIBUTIONS ON UNSUPERVISED LEARNING

Laplace, great mathematician of the Enlightment, once stated that "ignorance can be expressed as uniform priors" (cited in Gigerenzer & Hoffrage, 1995). This affirmation was motivated by his understanding of the world, which is basically made of biases. For instance, it is common knowledge that one is more likely to run into dogs than wolves in a city. This information is useful when a partially occluded animal, which might be either a dog or a wolf, is encountered downtown. In this circumstance, the most probably correct inference is that the animal is a dog. The same type of reasoning is sound when waking up at night and seeing a dog while the lights are off. Even though there is not enough visual information to identify with certainty the family's pet, it is rational to infer so, because the probability that another dog is in one's own room at night varies from slight to null. In the former example, the frequency of categories was used to correctly identify an unknown item while, in the latter, it was the frequency of exemplars.

Depending on the level of analysis, both these frequencies must be taken into account in order to achieve rational or optimal behavior (Kahneman, Slovic & Tversky, 1982). When facing uncertain information, the rational (normative) way of computing category membership is to employ Bayes’ theorem (Anderson, 1991; Oaksford & Chater, 1999), which uses prior odds to compute the *posterior probability* (viz. the updated probability of category membership after consideration of new information):

$$\begin{aligned} P(C_j \setminus O) &= \frac{P(C_j O)}{P(O)} \\ &= \frac{P(O \setminus C_j) P(C_j)}{\sum_i P(O \setminus C_i) P(C_i)} \end{aligned} \tag{1}$$

where C_j is category *j* and *O* is a new observation. In words, the probability of category membership, given a new observation, is dependent of its *a priori probability* and its inverse *conditional probability*. The resulting posterior probability is optimal according to the maximum likelihood criterion, which justifies the use of Bayes' theorem in belief networks (Pearl, 1988). These *Bayesian networks* are commonly employed as benchmarks in machine learning (Russell & Norvig, 2003).

Not only are base rates necessary to increase the performance of applications in computer science but cognitive models must also account for them. In fact, an increasing body of evidence is now showing that humans are using base rate information (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Körding & Wolpert, 2004; Wagman, 2003). For instance, it has been shown that presenting a problem in frequentist format, in opposition to probability (e.g., 10 / 100 vs. 0.1), elicits the use of base rates as well as Bayes' theorem in human participants (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). According to Gigerenzer and Hoffrage, the advantage of the frequentist format stems from the observation that frequency is a natural way to encode probabilities by simply maintaining counts of past experiences. Cosmides and Tooby further argued that using frequentist representations is adaptive and that the use of this type of encoding is a result of evolutionary processes.

Beside the possible phylogenic advantage of encoding frequency information

(Cosmides & Tooby, 1996), the effects of category and exemplar frequency have received much attention in cognitive psychology since the mid eighties (Erickson & Kruschke, 1998; Kruschke, 1996; Nosofsky, 1988, 1991; Nosofsky & Palmeri, 1997; Rips, 1989; Shin & Nosofsky, 1992). For instance, Rips (1989) has provided participants with histograms showing the frequency distributions of temperatures measured in January and July. Together, these two data sets formed a bimodal distribution in which middle temperatures were absent. After inspection of the histograms, the participants were asked to infer if new temperatures had been measured in January or July. The results of this experiment showed that the categorization judgments are sensitive to frequency: the participants classified the new temperatures in the same category as the most frequent alternative.

Kruschke (1996) has found a similar effect of category frequency by asking participants to use a group of symptoms to classify diseases. In a series of experiments, including various levels of cue-validity, the classification results reflected the prior probabilities of the diseases when the cues did not provide sufficient information to confidently classify the stimuli. Moreover, the participants were afterward asked to estimate the frequency of appearance of each disease. The odd resulting from the participants' guesses showed that they were aware of the bias toward a particular disease and have been consciously using this information.

Concerning the effect of exemplar frequency, it was mostly studied in the exemplarist framework by Nosofsky and his colleagues (Nosofsky, 1988, 1991; Nosofsky & Palmeri, 1997; Shin & Nosofsky, 1992). In Nosofsky (1988), typical and atypical members of the *pinkish* and *brownish* Munsell color categories were biased in order to verify: 1) the effect of exemplar frequency and, 2) the interaction between exemplar typicality and frequency. The results showed that more frequent exemplars were better categorized (and faster, see Nosofsky & Palmeri, 1997), notwithstanding typicality. Also, this advantage of biased stimuli spread to other category members which were similar. On the other hand, classification accuracy of stimuli in the opposite category, which were nevertheless similar to biased stimuli, was decreased. According to Anderson (1991), this (dis) advantage of the region surrounding biased exemplars can be simply explained by a shift of the category's center toward the biased stimuli. Regarding exemplar typicality, more frequent stimuli were judged more typical after training.

In another series of experiments, Nosofsky (1991) tested the effect of exemplar frequency on the recognition and categorization performances of Brunswick faces (Reed, 1973). The results have shown that frequency of presentation affected both recognition and categorization. However, the latter had a greater advantage than the former. Also, other experiments involving abstract polygons (Homa, Dunbar & Nohre, 1991) showed that the size of the frequency effect is not modulated by the size of the category (Shin & Nosofsky, 1992).

Erickson and Kruschke (1998) later found a much stronger result concerning the effect of exemplar frequency on categorization. In developing ATRIUM, a hybrid model which uses both an exemplarist (ALCOVE: Kruschke, 1991) and a rule module, Erickson and Kruschke created a categorization task in which some stimuli could be classified using a deterministic rule while others were exceptions that needed to be dealt with by the exemplarist module. The aim of this manipulation was to isolate the effect of each module, and it was presumed that frequency of exemplar would have an effect on exception stimuli but not on those covered by the rule. Surprisingly, frequency effects were found on both rule and exception stimuli. Moreover, the magnitude of the effect was the same.

Together, all the preceding results make a strong case for the presence of category and exemplar frequency effects on human performance: clearly, humans aren't "ignorant" (in the Laplacian sense). Are psychological models of human categorization able to account for all these results? Past arguments have been made about the inadequacy of backpropagation neural networks to plausibly model environmental feedback (for example, see Proulx & Hélie, 2005). As a result, several modelers fell back on unsupervised learning to train connectionist models (Anderson et al., 1977; Barlow, 1989; Bégin & Proulx, 1996; Carpenter & Grossberg, 1987; Grossberg, 1976_a, 1976_b; Kohonen, 1984; Proulx & Hélie, 2005; Rumelhart & Zipser, 1986). However, very little is known about the capacity of unsupervised artificial neural networks (ANNs) to learn environmental biases. Hence, the aim of the present paper is to test the "ignorance" of two popular families of such networks: recurrent associative memories (RAMs: e.g., Anderson et al., 1977; Hopfield, 1982) and competitive networks (e.g., Carpenter & Grossberg, 1987; Kohonen, 1984; Rumelhart & Zipser, 1986).

Unsupervised Neural Networks

Unsupervised ANNs have been extensively used in computer science (Bishop, 1995; Cichocki & Unbehaun, 1993; Russell & Norvig, 2003) and psychological modeling (Anderson et al., 1977; Carpenter & Grossberg, 1987; Grossberg, 1976_a, 1976_b; McClelland, 1998). In particular, competitive networks (e.g., Carpenter & Grossberg, 1987; Grossberg, 1976_a, 1976_b; Kohonen, 1984; Nowlan, 1989; Rumelhart & Zipser,

1986) and RAMs (e.g., Anderson et al., 1977; Bégin & Proulx, 1996; Chartier & Proulx, 2005; Hopfield, 1982) are the most popular families of unsupervised ANNs and their "ignorance" has never been tested directly. RAMs, which are usually trained using some variant of Hebbian learning (Kohonen, 1972), are known to minimize the following Lyapunov function (Cohen & Grossberg, 1983; Diamantaras & Kung, 1996; Golden, 1986):

$$E(\mathbf{X}) = -\tfrac{1}{2}\mathbf{x}^{\mathbf{T}}\mathbf{W}\mathbf{x} \tag{2}$$

where **x** is a stimulus vector and **W** is the weight matrix. After training, this energy function represents half the negative of the output's variance and, hence, RAMs trained using hebbian learning can be understood as unstable principal component analyzers (Diamantaras & Kung, 1996). However, no clear statement is made about RAMs' sensitivity to environmental biases.

Competitive networks can be separated in two distinct categories: hard competitive learners (e.g. Carpenter & Grossberg, 1987; Grossberg, 1976_{a}, 1976_{b}; Rumelhart & Zipser, 1986) and soft competitive learners (e.g. Kohonen, 1984; Nowlan, 1989). The former refers to networks in which only the winning unit's weights are updated (the winning unit is the closest to the stimulus shown according to some metric) while the latter refers to networks in which all weights are updated by an amount inversely related to their distance from the stimulus shown. The topic of this paper is restricted to hard competitive learners[1]. Competitive networks are known to maximize the overlap of stimuli and weights (Rumelhart & Zipser, 1986). In other words, a stable weight vector corresponds to the average of the stimuli that maximally activates it (Nowlan, 1989), and the network performs a k-means cluster analysis (Hastie, Tibshirani

& Friedman, 2001). Again, it is unclear whether this averaging accounts for environmental biases.

If RAMs and competitive learners turn out to be “ignorant” to the environmental biases, they would be poor alternatives to backpropagation neural networks for two reasons. First, ignorant models are unfit for engineering purposes because they are bounded to sub-optimality. Second, they would poorly reflect human performance, in which there is no apparent “ignorance” concerning these biases.

To summarize, whether the goal is to model human cognition or create an optimal ANN in a given application context, the chosen ANN must be able to absorb the environmental distribution in order to correctly fulfill its duty. Competitive neural networks and RAMs are used in both these contexts but they have never been tested as to their capacity to reflect environmental biases. This is precisely the aim of the present work.

Overview

In order to test the whole families of competitive networks (e.g. Carpenter & Grossberg, 1987; Grossberg, 1976_a, 1976_b; Rumelhart & Zipser, 1986) and RAMs (e.g. Anderson et al., 1977; Bégin & Proulx, 1996; Chartier & Proulx, 2005; Hopfield, 1982), two sets of simulations were performed. In the first set, the simulations were completed with the simplest network of each, in the simplest possible environment: an orthonormal world composed of two categories. Therefore, every stimulus was represented by a unit vector and the categories were orthogonal. It is important to note that, if the networks are unable to estimate prior odds in this simple environment, it follows that they are unable to estimate those odds in any other environments.

In the second set of simulations, complexity was increased by using more sophisticated networks in more complex environments. In this case, correlated patterns from three different categories were used. These simulations were performed to test whether the behavior of these families of ANNs remained analogous in situations closer to the real world.

Obviously, the more sophisticated networks could have been used in both simulation sets. However, the choice of using simpler networks in the simple environment was motivated by: 1) simpler networks are usually linear, which allows in-depth analyses of their performance and, 2) one of the goals of this paper was to test the learning capacity of the whole families of networks. Hence, if the complex representative of one family of networks happens to be “ignorant”, this incapacity might result from some superfluous axiom. However, the simple networks used in the first simulation set had the minimal set of axioms needed to be a member of the RAMs’ or the competitive networks’ families. Therefore, the incapacity of one of these networks to reflect environmental biases would make a strong case about an important flaw present in all the members of its family.

Simulation set 1: The orthonormal world

Stimuli

The stimuli used for the simulations are shown in Fig. 1. As seen, the stimuli were bipolar vectors composed of eight units. Because each stimulus was represented by a unit vector, the white and black squares were coded as $\pm 8^{-1/2}$. The dimensionality of the stimuli was chosen to be low in order to test the models in a simple situation: if the models are unable to estimate frequencies in low dimensionality, it is useless to test them in higher dimensional spaces.

Insert Fig. 1 about here

As shown in Fig. 1, the first three stimuli were arbitrarily labeled as category “A” and the remaining as category “B”. The within-category correlation was constant and equal to 0.5. The between-category correlation was null, because the categories were orthogonal.

Models: The two ANNs were a simple Hebbian learning RAM (Anderson et al., 1977) and the simplest competitive network (Rumelhart & Zipser, 1986). As for all ANNs, they can be entirely described by their architecture, transfer function, and learning rule.

RAM

The architecture of the RAM used in this set of simulations is shown in Fig. 2a. In the present case, the network was composed of eight units. The transmission rule is described by:

$$\mathbf{x}_{[t+1]} = f(\mathbf{W}\mathbf{x}_{[t]} + \phi\mathbf{x}_{[t]}) \tag{3}$$

where $\mathbf{x}_{[t]}$ is the stimulus at time *t*, **W** is the weight matrix and ϕ is a restraining parameter (Bégin & Proulx, 1996). When ϕ is set to null, this transmission rule is identical to the one proposed by Hopfield (1982), and when it is set to one, it is identical to the rule proposed by Anderson et al. (1977). It is important to note that the value of this parameter does not affect the emergent properties of the model: when used with iterative Hebbian learning (Eq. 5), ϕ can take any non-null value and only the speed of convergence is affected. Hence, for simplicity, ϕ

was set to one in the present simulations (as hinted by the block diagram). The transfer function was a saturation limiter described by:

$$f(x) = \begin{cases} +1, x > 1 \\ x, -1 \leq x \leq 1 \\ -1, x < -1 \end{cases} \quad (4)$$

where x is the activation of a given unit. This saturation function was used in the *Brain-State-in-a-Box* (Anderson et al., 1977).

Insert Fig. 2 about here

The learning rule was a simple hebbian function (Kohonen, 1972) described by:

$$\mathbf{W}_{[k]} = \mathbf{W}_{[k-1]} + \eta(\mathbf{x}_{[p]}\mathbf{x}_{[p]}^{\mathsf{T}}) \quad (5)$$

where $\mathbf{W}_{[k]}$ is the weight matrix at the k^{th} trial, the second term is the weight change, η is a learning parameter and p is the number of iterations in the network prior to learning. As suggested by Anderson and his colleagues (1977), p was set to 7 and η to 0.0001.

Competitive neural network

The simple competitive model proposed by Rumelhart and Zipser (1986) was used in this simulation set. Its architecture is that of a standard feedforward network composed of two layers with inhibitive lateral connections in the output layer. One such architecture is shown in Fig. 2b. In the present work, the network was composed of eight input units and two output units (because there are two categories). The transmission in the network is described by:

$$WinningUnit = \underset{i}{Min}\left\|\mathbf{w}_{\mathbf{i}} - \mathbf{x}\right\|_2 \tag{6}$$

where $\mathbf{w}_\mathbf{i}$ is the i^{th} vector of the weight matrix, $\mathbf{x}$ is the stimulus-vector and $\|\bullet\|_2$ is the Euclidean distance (L2-Norm). The learning rule was applied only to the winning unit. The weight update is described by:

$$\mathbf{w}_{i[k]} = \mathbf{w}_{i[k-1]} + \eta(\mathbf{x} - \mathbf{w}_{i[k-1]}) \tag{7}$$

where $\mathbf{w}_{\text{i}[k]}$ is the weight vector of the winning unit (i) at the k^{th} trial, $\mathbf{x}$ is the stimulus, and η is a learning parameter. The second term represents the weight update. In the present, η was set to 0.01.

Simulations

The aim of the present work was to test the effect of a change in the environmental distribution on the behavior of unsupervised ANNs. To fully apprehend the problem of environmental biases, environmental distributions were chosen in order to vary the frequency of exemplars and categories independently. This was accomplished by using four different distributions which are shown in Fig. 3. In the first condition, the networks were trained using a uniform distribution (Fig. 3a): As a result, all categories and all exemplars were equally likely (control condition). The second condition used a bimodal distribution composed of two Gaussians with respective means of 2 and 5, and a common standard deviation of 0.5 (Fig. 3b). Therefore, both categories appeared as often but exemplar two (from category “A”) and exemplar five (from category “B”) were about four times more likely than the others. In the third condition, the networks were trained using a step distribution (Fig. 3c): items from category “A” were more probable than those from category “B”. More precisely, the ratio of “A” to “B” stimuli was 5:4. Within each category, all exemplars were equiprobable. Finally, the fourth condition used an exponential distribution with a mean of 3 (Fig. 3d). Therefore, category “A” was more probable than “B” (about 3:1), and exemplar one was more probable than exemplar two, which was more probable than three, etc.

Insert Fig. 3 about here

A different competitive network was trained in each condition for 500 trials. After this training, each network was tested by presenting 500 random vectors composed of real numbers from the interval [-1, 1]. The aim of this Monte Carlo simulation was to estimate the content of the learned categories by counting the number of random vectors classified in each category. The same simulation methodology was used to train and test the RAM. Simulations and random number generation were conducted using Mathematica (Wolfram, 1996).

Results

Training

First, both networks were able to recall perfectly every training exemplar in each condition, which indicated that they were properly trained. Second, because the categories were orthogonal, the eigenvectors of the weight matrices were the energy minima in the RAMs (Diamantaras & Kung, 1996). Therefore, we proceeded with a spectral decomposition. Fig. 4 shows the eigenvalues of the weight matrices in decreasing order. As seen, the amplitude of the eigenvalues varied as a function of the frequency of the categories. For example, panel (a) and (b) show conditions in which both categories were equally likely (Uniform and Bimodal, respectively): hence, both developed eigenvalues were almost identical (mean difference = 1.89 x 10^{-3}). However, in conditions where category "A" was more likely than category "B", the eigenvalue associated with category "A" was bigger than the eigenvalue associated with category "B". For instance, in the condition where the RAM was trained with the step distribution (panel c), the difference between the first and the second eigenvalue was three times bigger than those from conditions in which the categories were equally likely; this difference became twelve times bigger when trained with the exponential distribution (panel d).

Insert Fig. 4 about here

Fig. 5 shows the eigenvectors corresponding to each category developed by the RAMs (after convergence). As seen, the position of the eigenvectors was affected by the relative intra-categorical frequency of exemplars. For instance, Fig. 5 shows that the eigenvectors from the uniform condition (a) were identical to those from the step condition (c). However, eigenvectors from the uniform condition differed from those from the bimodal (b) and exponential (d) conditions. This difference reflects a rotation of the eigenvectors toward stimuli that were more frequent. The bimodal condition clearly illustrates this phenomenon (panel b). Comparing its eigenvectors with the stimuli (Fig. 1) brings forward the similarity between the "A" eigenvector and stimulus two (remember that stimulus two was four times more likely than other "A"s). The same observation is made when comparing the "B" eigenvector with stimulus five. Because showing stimulus one more frequently saturated the first eigenvector, the exponential condition (panel d) reflects the same phenomenon in a less obvious way.

Insert Fig. 5 about here

In the introduction, it was asserted that competitive networks are maximizing the overlap of stimuli and weight vectors (Rumelhart & Zipser, 1986). Therefore, the learned categories should be apparent in the weight matrices. Fig. 6 shows these weights after saturation in a comparator[2]. As shown, weights from the first three panels were identical to panel (a) to (c) of Fig. 5. This similarity confirms that both networks learned the same categories in these three conditions (Uniform, Bimodal, and Step). However, there was a difference in the exponential condition between the RAM (Fig. 5d) and the competitive network (Fig. 6d). As seen, the competitive network's weights were not moved towards the first and fourth stimuli from category "A" and "B" respectively. Yet, this rotation towards more frequent stimuli was present in the bimodal condition (Fig. 6b). This difference reflects a certain coarseness of the network according to changes in frequency of exemplars: in the bimodal case, exemplar two and five were four times more likely than the remaining; in the exponential case, this ratio was only 7:5 between the first and second stimuli, and 2:1 between the first and third stimuli (similar differences were computed between stimulus four and five and stimulus four and six). These smaller odd ratios were insufficient to attract the weight vectors toward the first and fourth stimuli.

Insert Fig. 6 about here

Test

Table 1 shows classification results of random vectors in all conditions for each network. As seen, the RAMs' density estimations did not significantly differed from the training distributions according to a χ^2 test. However, the competitive networks were unable to estimate the correct densities in both the step and exponential environments. More specifically, the competitive networks' density estimations never differed from a uniform environment (all $\chi^2(1) < 6.728, p > .001$). On the other hand, density estimations of the RAMs in the step and exponential conditions significantly differed from a uniform distribution (both $\chi^2(1) > 23.13, p < .001$).

Insert Table 1 about here

Discussion

The aim of this set of simulations was to test the effect of the environmental distribution on the simplest representatives of two important classes of unsupervised ANNs: competitive networks (e.g. Carpenter & Grossberg, 1987; Grossberg, 1976_a, 1976_b; Rumelhart & Zipser, 1986) and RAMs (e.g. Anderson et al., 1977; Bégin & Proulx, 1996; Chartier & Proulx, 2005; Hopfield, 1982). Our simulations used various types of random-generating functions in order to test the effect of the frequency of exemplars and categories in simple, orthonormal environments. Tests involving random vectors have shown that only the RAMs' performances were affected by this manipulation. Moreover, the effects of exemplar and category frequency were independent: the former was reflected by the position of the eigenvectors of the weight matrices and the latter by the magnitude of the eigenvalues. Hence, a more frequent category at training resulted in more random vectors being categorized as members of that category, which is consistent with empirical data in category learning (Kruschke, 1996; Rips, 1989). Also, in order to better model the possible discrepancy between objective frequencies and human estimated frequencies, Anderson and his colleagues (1977) have shown that adding a memory efficiency parameter in the learning rule (first term of Eq. 5) allows for (under / over) estimation of category frequency. Concerning the effect of stimulus frequency, a change in the relative frequency of the exemplars was reflected by a displacement of the attractors that determined which random vectors were classified in which category. This observation is in line with Anderson's explanation of the stimulus frequency effect (Erickson & Kruschke, 1998; Nosofsky, 1988, 1991; Nosofsky & Palmeri, 1997; Shin & Nosofsky, 1992): the center of the category is moved toward biased exemplars (Anderson, 1991). This effect was also present in the position of the weight vectors of the competitive networks. Nevertheless, the environmental biases did not affect these networks' performance: random vectors were categorized as members of both categories half of the time. Clearly, the competitive networks' results did not reflect environmental biases, thus revealing their incapacity to model human categorization data in tasks where category frequency is varied.

Simulation set 2: Increased complexity

The aim of the present simulation set was to test the behavior of unsupervised neural networks in more complex environments, which included more stimuli pertaining to more than two categories[3]. The patterns were highly dimensional and correlated.

Stimuli

The stimuli used for the simulations were thirty hand-written digits drawn using a 16 × 16 grid (shown in Fig. 7). Each stimulus was coded as a bipolar 256-units vector. These stimuli were chosen because they represent a wide range of correlations. The within-category correlations varied between 0.74 and 0.95 while the between-category correlations varied between 0.22 and 0.42.

Insert Fig. 7 about here

Models

To correctly classify the stimuli shown in Fig. 7, a novelty detector must be added to the RAM and the competitive network. The notion of novelty detection (or vigilance) was introduced in Grossberg (1976$_a$), and further developed in Carpenter and Grossberg (1987), to solve the stability / plasticity dilemma: when new stimuli are introduced after learning has started, the system must be able to learn them without forgetting what is already known. Grossberg proposed to achieve this by adding a vigilance module, which controls the granularity of the categorization. Hence, if the new stimuli are sufficiently different from existing classes, a new category is created, leaving the previous knowledge untouched. The size of the classes is defined by the value given to the vigilance parameter, a small value resulting in few categories while a high value in rote learning.

In the present case, the two ANNs used were state-of-the-art representatives of the competitive networks' and RAMs' families which can classify bipolar vectors. As a result, ART1 (Carpenter & Grossberg, 1987) was used to represent the competitive network's family and NDRAM (Chartier, 2004; Chartier & Proulx, 2005) was used to represent the RAMs'.

NDRAM

NDRAM is a non-linear RAM which can categorize grey-level correlated patterns with few spurious states (Chartier & Proulx, 2005). The ability to learn continuous stimuli is new in RAMs and stems from the non-linearity of the transfer function and its inclusion into the learning rule (for details, see Appendix A). The small quantity of spurious states greatly improves the network's performance and follows from the learning rule, which ensures that NDRAM's eigenvalues are converging toward an unequal spectrum. However, in NDRAM, there are as many memory traces as there are distinct stimuli. Hence, a vigilance parameter for novelty detection must be added to achieve categorization (Chartier, 2004; Chartier & Proulx, 1999).

NDRAM's architecture is shown in Fig. 8a. As seen, the main structure is that of a usual associative memory, but an external novelty detector was added. This module computes the correlation between the input and the output. If this correlation is higher than a predetermined criterion (ρ), the module returns a balanced average of the input and the output ($\bar{\mathbf{x}}$) for learning. Otherwise, the input is used for learning ($\bar{\mathbf{x}} = \mathbf{x}_{[0]}$).

Insert Fig. 8 about here

In the present case, the network was composed of 256 units (N). Following Chartier and his colleagues' analyses (Chartier, 2004; Chartier & Proulx, 2005), η and δ were set to 0.001 and 0.4 (which ensures that NDRAM's attractors are steady), ζ and μ were set to 0.9999 and 0.01, and the vigilance parameter (ρ) and number of iterations (p) were set to 0.7 and 10 respectively.

ART1

ART1 (Carpenter & Grossberg, 1987) is a biologically inspired competitive network based on two main ideas: resonance (Grossberg, 1976_b) and novelty detection (vigilance; Grossberg, 1976_a). Resonance is a state of equilibrium reached when an output is able to reconstruct the input that generated it. An ART1 network only learns when this equilibrium is reached. The second idea is a fuzzy criterion that permits resonance to take place: instead of requiring that the output's reconstruction be identical to the input to qualify as an equilibrium state, a predetermined level of similarity must be reached. If the reconstruction of the output is sufficiently similar to the input, the stimulus is a member of the output's category. Otherwise, the input is a member of another category (for details, see Appendix B).

The architecture of ART1 is shown in Fig. 8b. As shown, there are two layers of units and the network is composed of two distinct sets of weights: one from F1 to F2 ($\mathbf{W_{bu}}$, for bottom-up) and another from F2 to F1 ($\mathbf{W_{td}}$, for top-down). The right part of the network is the novelty detector.

In the present case, the input layer (F1) was composed of 256 units and the output layer (F2) was composed of three units (because there are three categories). Following Freeman (1994), $a = 1$, $b = 1.5$, $c = 5$ and $d = 0.9$. It is noted that these parameter values satisfy all constraints listed in Carpenter & Grossberg (1987). To further meet these constraints, L was set to 45 (because of the high dimensionality of the network) and ρ was set to 0.4.

Simulations

As in the first simulation set, four environments were chosen in order to manipulate independently the biases of stimuli and categories. The chosen distributions are shown in Fig. 9. Again, the control condition was described by a uniform distribution, this time ranging from 1 to 30 (panel a). The second condition was a multimodal distribution with three modes (stimuli 5, 15, and 25). This environment was created by a mixture of three Gaussian distributions with means corresponding to modes and a common standard deviation of two (panel b). The step distribution is shown in Fig. 9c. In this condition, the ratio of the first category (“7”) to the second (“0”) was 7:5, and the one from first to third (“1”) was 7:4. Finally, panel (d) shows an exponential distribution with a mean of 25. In this condition, the ratio of the first category to the second was 3:2, and the ratio from the first to third was 9:4. As in the first simulations, the uniform distribution constituted an unbiased environment, the multimodal distribution formed a world in which every category was equally likely but some exemplars were biased, the step distribution resulted in biased categories while keeping within-category exemplar frequency constant, and the exponential condition resulted in biased categories and exemplars.

Insert Fig. 9 about here

A different ART1 was trained in each condition for 10 000 trials. After this training, each network was tested with 500 random vectors composed of real numbers from the interval [-1, 1] to evaluate the content of the learned categories. The same simulation methodology was used to train and test NDRAM. Simulations and random number generation were conducted using Mathematica (Wolfram, 1996).

Results

Training

As in the first simulation set, both networks were able to recall perfectly every training exemplar in each condition, indicating that the learning phase was successful. In the first simulation set, the presence of categorical biases was hinted by the eigenvalue spectrum of the RAMs’ weight matrices. Therefore, a spectral decomposition was performed on NDRAM’s weight matrices. Fig. 10 shows the eigenvalue spectrum developed in each condition. Unlike the first simulation set, the eigenvalue spectrum was not affected by the manipulations. The between-condition differences were of the order of 1×10^{-5}.

However, in the present case, the categories were composed of correlated patterns: as a result, categories were not equivalent with eigenvectors but with converged linear combinations of the eigenvectors. Therefore, each eigenvalue cannot be directly interpreted as associated to a given category; categorical biases might still be present in the test phase.

Insert Fig. 10 about here

In the first simulation set, the eigenvectors were drawn toward biased exemplars. Chartier (2004) has conjectured that NDRAM's attractors are the theoretical means of the categories. Therefore, this effect should still be present in NDRAM's categories. Fig. 11 shows NDRAM's categories in each condition. As predicted by Chartier's previous work, the attractors corresponding to each category were the theoretical means and were easily identifiable. However, because the within-category correlations were high (mean within-category $\underline{r} = 0.87$), the correlations between the theoretical means of the conditions were also high (mean $\underline{r} > 0.9$). Therefore, this attractor shift, while present, is difficult to appreciate visually.

Insert Fig. 11 about here

In the first simulation set, it was shown that the competitive networks' weights are affected by stimulus frequency (Fig. 6). However, Eq. B4 (from Appendix B) suggests that $\mathbf{W_{bu}}$ is equal to the intersection of the stimuli in a given category (not the mean, as in the first set of simulations): hence, the effect of exemplar frequency is unlikely to be present. The developed categories are shown in Fig. 12. As seen, the categories are easily identifiable and no grey-levels were present: each entry in the weight vector was either equal to $\dfrac{L}{L-1+\sum_j \mathbf{F1}'_j}$ or zero. Also, the categories did not seem to be affected by the frequency of exemplars.

Insert Fig. 12 about here

Test

Table 2 shows the classification results of random vectors in each condition for all networks. As seen, NDRAM (Chartier, 2004; Chartier & Proulx, 2005) displayed the same pattern of results as the RAM in the first simulation set: its density estimations did not significantly differed from the training distributions according to a χ^2 test. However, like the competitive network in the first simulation set, ART1 (Carpenter & Grossberg, 1987) was unable to estimate the correct densities. More surprisingly, ART1 did not display uniform priors in every condition (all $\chi^2(2) > 367.4$, $p > .001$): it was biased toward the "1" category, even in conditions where it was the less frequent. This bias is explained by a preference of the network for categories in which the intersection of the members has more position filled with non-null values. More precisely, the expectation of the activation of a given unit is[4]:

$$E(a) = \frac{L}{L-1+m}\left(\frac{1}{2}\right)^N \sum_{n=0}^{N}\binom{N}{n}\sum_{i=0}^{n}\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}i \tag{8}$$

where N is the cardinality of the stimulus, n is the number of positions containing a non-null value in the random vector, and m is the number of non-null values in the weight vector. This function is non-linear and increasing for all m (further details are given in

Appendix B). In the present case, $m = \{59, 50, 32\}$ for categories “1”, “0” and “7” respectively. Therefore, in addition to being insensitive to priors, ART1 is biased toward categories in which members are sharing more features. This bias is not viable, neither in psychological models nor in machine learning.

Insert Table 2 about here

Discussion

The aim of this set of simulations was to test the effect of the environmental distribution on sophisticated representatives of two important classes of unsupervised ANNs: competitive networks (e.g. Carpenter & Grossberg, 1987; Grossberg, 1976_a; Grossberg, 1976_b; Rumelhart & Zipser, 1986) and RAMs (e.g. Anderson et al., 1977; Bégin & Proulx, 1996; Chartier & Proulx, 2005; Hopfield, 1982). The simulations used various types of random-generating functions in order to test the effect of the frequency of exemplars and categories in environments containing more than two categories composed of correlated patterns. Tests involving random vectors as stimuli have shown that only NDRAM’s (Chartier, 2004; Chartier & Proulx, 2005) performance was affected by this manipulation. Moreover, the exemplar frequency effect was independent from the category frequency effect. The former was reflected by the position of the attractors, which were equal to the theoretical means of the categories and the latter was seen in the results of the Monte Carlo simulations (Table 2): A more frequent category at training resulted in more random vectors categorized as members of that category, which is in line with empirical data (Kruschke, 1996; Rips, 1989). A change in the relative frequency of the exemplar was reflected by a displacement of the attractors that determined which random vectors are classified in which categories, again consistent with human data (Erickson & Kruschke, 1998; Nosofsky, 1988, 1991; Nosofsky & Palmeri, 1997; Shin & Nosofsky, 1992). Neither of these effects was present in ART1's (Carpenter & Grossberg, 1987) weight matrices. Moreover, ART1 is biased toward categories in which the cardinality of the set, composed of the intersection of all the category members, is the highest. This bias is not desirable in a model because there is no a priori reason to believe that categories composed of members that strongly overlap should be privileged. On the contrary, human data suggests that sparse categories elicit the inclusion of more distorted patterns (Posner & Keele, 1968).

General Discussion

The aim of this paper was to test the "ignorance" (in the Laplacian sense) of two important classes of unsupervised ANNs: competitive networks (e.g. Carpenter & Grossberg, 1987; Grossberg, 1976$_a$; Grossberg, 1976$_b$; Rumelhart & Zipser, 1986) and RAMs (e.g. Anderson et al., 1977; Bégin & Proulx, 1996; Chartier & Proulx, 2005; Hopfield, 1982). The networks' "ignorance" was assessed by throwing random vectors in their weight space developed after training. Two simulation sets were used to train simple and complex networks in different environments, which varied the frequency of exemplars and categories independently.

First, all the simulations results' showed that the networks were able to correctly recall the training sets. Also, the categories were all visible in the weight matrices, which confirmed that the categories had been correctly learned. However, while RAMs were able to correctly estimate the training densities, competitive networks were unable to do so. The problem with hard competitive learners is their decision function (Max or Min).

For instance, the output layer of ART1 (Carpenter & Grossberg, 1987) in the second simulation set was three-dimensional. After a stimulus passes through the weight matrix, all three output units are activated, which can be represented by a vector in the three-dimensional output space. However, the output function chooses the maximally activated value and the remaining are shut down. Therefore, the output space is compressed into a one-dimensional space which is the dominating axis of the three-dimensional output space. All other information is lost, including rotations and dilatations of the weight vectors (due to exemplar frequencies) in the other directions.

These decision functions (Min, Max) are at the core of hard competitive learning and are a necessity for this kind of learning. This flaw thus affects the whole class of hard competitive learners, and frequency effects present in human data (Erickson & Kruschke, 1998; Kruschke, 1996; Nosofsky, 1988, 1991; Nosofsky & Palmeri, 1997; Rips, 1989; Shin & Nosofsky, 1992) cannot be explained unless the competition criterion is released (for a complete analysis of soft competitive learning, see Nowlan, 1989).

As a result, these findings have very important implications for ANN's modeling. First, because these tests were conducted on "generic" and sophisticated examples of competitive learners and RAMs, the properties brought forward by the simulations can be generalized to every ANNs based on those same axioms (hard competition vs. generalized Hebbian learning). Therefore, this work highlights an important flaw of hard competitive networks: they are unable to use the information present in the environment to optimize their behavior. This shortcoming limits the performance, and thus the applicability, of hard competitive networks in computer science. Moreover, it questions these models' adequacy as an explanation of human behavior.

On the other hand, RAMs may be useful models in computer science as well as psychology (Anderson et al., 1977; Bégin & Proulx, 1996; Chartier & Proulx, 2005). Their pattern completion capabilities as well as their ability to optimize recall by using environmental biases widen their field of application. However, in order to fully asses their usefulness, further study is needed with grey-level patterns (Chartier, 2004; Chartier & Proulx, 2005) and heteroassociative, BAM learning (Chartier & Boukadoum, in press; Kosko, 1988).

References

Anderson, J.R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413-451.

Barlow, H.B. (1989). Unsupervised learning. *Neural Computation*, *1*, 295-311.

Bégin, J. & Proulx, R. (1996). Categorization in unsupervised neural networks: The Eidos model. *IEEE Transactions on Neural Networks*, *7*, 147-154.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Carpenter, G.A. & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54-115.

Chartier, S. (2004). *Un Nouveau Modèle de Réseaux de Neurones Artificiels à Attracteurs dans le Cadre de la Catégorisation Autonome*. Ph.D. Thesis, Department of Psychology: Université du Québec À Montréal.

Chartier, S. & Boukadoum, M. (in press). A bidirectional heteroassociative memory for binary and grey-level patterns. *IEEE Transactions on Neural Networks*.

Chartier, S. & Proulx, R. (2005). NDRAM: Nonlinear dynamic recurrent associative memory for learning bipolar and non-bipolar correlated patterns. *IEEE Transactions on Neural Networks*, *16*, 1393-1400.

Chartier, S. & Proulx, R. (1999). A self-scaling procedure in unsupervised correlational neural networks. *Proceeding of the International Joint Conference on Neural Networks* (pp. 1092-1096). Washington.

Cichocki, A. & Unbehaun, R. (1993). *Neural Networks for Optimization and Signal Processing*. New York: Wiley.

Cohen, M.A. & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*, 815-826.

Cosmides, L. & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1-73.

Diamantaras, K.I. & Kung, S.Y. (1996). *Principal Component Neural Networks: Theory and Applications*. Toronto: John Wiley & Sons.

Erickson, M.A. & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107-140.

Freeman, J.A. (1994). *Simulating Neural Networks with Mathematica*. New York: Addison-Wesley Publishing Company.

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704.

Golden, R.M. (1986). The “Brain-State-in-a-Box” neural model is a gradient descent algorithm. *Journal of Mathematical Psychology*, *30*, 73-80.

Grossberg, S. (1976_a). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121-134.

Grossberg, S. (1976$_b$). Adaptive pattern classification and universal recoding: II. Feedback, expectation, olfaction, illusions. *Biological Cybernetics*, *23*, 187-202.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 444-458.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554-2558.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers C*, *21*, 353-359.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. New York: Springer-Verlag.

Körding, K.P. & Wolpert, D.M. (2004). Bayesian integration in sensorymotor learning. *Nature*, *427*, 244-247.

Kosko, B. (1988). Bidirectional associative memories. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*, 49-60.

Krushcke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Kruschke, J.K. (1996). Base rates in category learning. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *22*, 3-26.

McClelland, J.L. (1998). Connectionist models and Bayesian inference. *In* M. Oaksford & N. Chater (Eds.) *Rational Model of Cognition* (pp. 21-53). New York: Oxford University Press.

Nosofsky, R.M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 54-65.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 3-27.

Nosofsky, R.M. & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.

Nowlan, S.J. (1989). Maximum likelihood competitive learning. *In* D.S. Touretsky (Ed.) *Advances in Neural Information Processing Systems* (pp. 574-582). San Mateo, CA: Morgan Kaufmann.

Oaksford, M. & Chater, N. (1999). *Rational Models of Cognition*. Oxford, UK: Oxford University Press.

Palm, G. (1982). *Neural Assemblies: An Alternative Approach*. New York: Springer-Verlag.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers Inc.

Posner, M.I. & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.

Proulx, R. & Hélie, S. (2005). Adaptive categorization and neural networks. *In* C.

Lefebvre & H. Cohen (Eds.) *Handbook of Categorization in Cognitive Science* (pp. 793-815). Amsterdam: Elsevier.

Reed, S.K. (1973). Perceptual vs. conceptual categorization. *Memory & Cognition*, *1*, 157-163.

Rips, L.J. (1989). Similarity, typicality, and categorization. *In* S. Vosniadou & A. Ortony (Eds.) *Similarity & Analogical Reasoning* (pp. 21-59). Cambridge, MA: Cambridge University Press.

Ross, S.M. (1998). *A First Course in Probability. Fifth Edition*. Upper Saddle River, NJ: Prentice-Hall.

Rumelhart, D.E. & Zipser, D. (1986). Feature discovery by competitive learning. *In* D.E. Rumelhart & J.L. McClelland (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (pp. 151-193). Cambridge: MIT Press.

Russell, S. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach. Second Edition*. Upper Saddle River, NJ: Prentice-Hall.

Shin, H.J. & Nosofsky, R.M. (1992). Similarity scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, *121*, 278-304.

Wagman, M. (2003). *Reasoning Processes in Humans and Computers: Theory and Research in Psychology and Artificial Intelligence*. Westport, CT: Praeger Publishers.

Wolfram, S. (1996). *The Mathematica Book*. New York: Cambridge University Press.

Acknowledgments

Part of this text was presented at the *26th Annual Meeting of the Société Québécoise pour la Recherche en Psychologie*, Montreal, Canada (2003). This research is part of the doctoral requirements of the first author and was supported by scholarships from *Le fonds québecois de la recherche sur la nature et les technologies* and the *Natural Sciences and Engineering Research Council of Canada* granted to the first author. The authors would like to thank Gyslain Giguère, Denis Cousineau, and Guy L. Lacroix for discussions and comments about the present research. Also, we would like to thank Dr. Axel Cleeremans, Dr. Lael Schooler, and an anonymous reviewer for their useful comments on an earlier draft.

Requests for reprints should be addressed to Sébastien Hélie, Laboratoire d'Étude en Intelligences Naturelle et Artificielle, Département de psychologie, Université du Québec À Montréal, C. P. 8888, Succ. Centre-ville, Montréal (Québec) H3C 3P8, CANADA, or using e-mail at Helie.Sebastien@courrier.uqam.ca.

Footnotes

[1] In the remaining, hard competitive learning is simply referred as competitive learning unless otherwise specified.

[2] Cichocki and Unbehauen (1993) defined the comparator as: $x = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$

[3] It is noted that there exist a qualitative difference between worlds composed of two categories and worlds composed of more than two. In the former case, a single category needs to be learned and the other can be defined as its complement. However, when more than two categories are present, increasing their number results in a quantitative difference.

[4] In the present, we adopt the standard convention that if $i < 0$, $n < 0$ or $n - i < 0$, $\binom{n}{i} = 0$ (Ross, 1998).

Appendix A

This Appendix presents the technical details of NDRAM (Chartier & Proulx, 2005) augmented with a novelty detector (Chartier, 2004). Its architecture is shown in Fig. 8a. This simplified architecture follows from NDRAM's transmission rule, in which ϕ (from Eq. 3) was set to zero. The resulting transmission rule is described by:

$$\mathbf{x}_{[\mathbf{t+1}]} = g(\mathbf{W}\mathbf{x}_{[\mathbf{t}]}) \tag{A1}$$

where $\mathbf{W}$ is the weight matrix and $\mathbf{x}_{[\mathbf{t}]}$ is the stimulus at time t. Most of the network's new behaviors arise from the non-linearity of the transfer function, which is described by:

$$g(a) = \begin{cases} +1, \; a > 1 \\ (\delta+1)a - \delta\, a^3, \;\; -1 \leq a \leq 1 \\ -1, \; a < -1 \end{cases} \tag{A2}$$

where a is the activation of a given unit and $0 < \delta < 2$ is a general transmission parameter (Chartier & Proulx, 2005). The learning rule is a simple hebbian function (Kohonen, 1972) with a correction term (anti-hebbian; Bégin & Proulx, 1996; Chartier & Proulx, 2005; Palm, 1982; Proulx & Hélie, 2005):

$$\mathbf{W}_{[\mathbf{k}]} = \zeta\mathbf{W}_{[\mathbf{k-1}]} + \eta(\overline{\mathbf{x}}\overline{\mathbf{x}}^{\mathbf{T}} - \mathbf{x}_{[\mathbf{p}]}\mathbf{x}_{[\mathbf{p}]}{}^{\mathbf{T}}) \tag{A3}$$

where the second term is the weight change, η is a learning parameter ($\eta < \frac{1}{2(1-2\delta)N}$, $\delta \neq 1/2$), p is the number of iterations in the network prior to learning, $0 << \zeta \leq 1$ is a general memory efficiency parameter, and $\overline{\mathbf{x}}$ is the output of the vigilance module.

The vigilance module is mathematically described by (Chartier, 2004):

$$\overline{\mathbf{x}} = \frac{z(\mu\mathbf{x}_{[\mathbf{0}]} + \mathbf{x}_{[\mathbf{p}]})}{1+\mu z}\mathbf{x}_{[\mathbf{0}]}(1-z) \tag{A4}$$

where $\mathbf{x}_{[\mathbf{0}]}$ is the stimulus, μ is a parameter which quantifies the effect of the initial stimulus in $\overline{\mathbf{x}}$, and *z* is defined by:

$$z = \begin{cases} 1, \dfrac{\mathbf{x}_{[\mathbf{0}]}\mathbf{x}_{[\mathbf{p}]}^{\mathbf{T}}}{\left\|\mathbf{x}_{[\mathbf{0}]}\right\|_{\mathbf{2}}\left\|\mathbf{x}_{[\mathbf{p}]}\right\|_{\mathbf{2}}} > \rho \\ 0, \dfrac{\mathbf{x}_{[\mathbf{0}]}\mathbf{x}_{[\mathbf{p}]}^{\mathbf{T}}}{\left\|\mathbf{x}_{[\mathbf{0}]}\right\|_{\mathbf{2}}\left\|\mathbf{x}_{[\mathbf{p}]}\right\|_{\mathbf{2}}} \leq \rho \end{cases} \tag{A5}$$

where *ρ* is the vigilance parameter.

Appendix B

Model

This part of the Appendix presents the equations defining the ART1 model. More details can be found in (Carpenter & Grossberg, 1987). The architecture of ART1 is shown in Fig. 8b. The output of the first layer (F1) is equal to the following:

$$\mathbf{F1} = \begin{cases} 1, & \dfrac{\mathbf{x}}{1 + a(\mathbf{x} + \mathbf{b}) + \mathbf{c}} > 0 \\ 0, & \dfrac{\mathbf{x}}{1 + a(\mathbf{x} + \mathbf{b}) + \mathbf{c}} \leq 0 \end{cases} \tag{B1}$$

where **x** is the chosen stimulus, and *a*, **b**, **c** are free parameters. This output is sent to the second layer (F2):

$$\mathbf{F2} = \begin{cases} 1, & \mathbf{F2}_i = Max(\mathbf{W_{bu}F1}) \\ 0, & \mathbf{F2}_i \neq Max(\mathbf{W_{bu}F1}) \end{cases} \tag{B2}$$

where $\mathbf{W_{bu}}$ is the bottom-up weight matrix from F1 to F2. The output vector (**F2**) is sent back to the input layer for comparison with **F1**. The reconstruction of the input is:

$$\mathbf{F1}' = \begin{cases} 1, & \dfrac{\mathbf{x} + d\mathbf{W_{td}F2} - b}{1 + a(\mathbf{x} + d\mathbf{W_{td}F2}) + c} > 0 \\ 0, & \dfrac{\mathbf{x} + d\mathbf{W_{td}F2} - b}{1 + a(\mathbf{x} + d\mathbf{W_{td}F2}) + c} \leq 0 \end{cases} \tag{B3}$$

where $\mathbf{W_{td}}$ is the top-down weight matrix connecting F2 to F1 and *d* is a free parameter. If $(||\mathbf{F1'}||_2 / ||\mathbf{F1}||_2)^{1/2} > \rho$, the input is recognized as a member of the output's category and the network is in a state of resonance. Else, the second highest output of Eq. B2 is set to one and the transmission is repeated until resonance is achieved. If the input is rejected by all known categories, resonance occurs with a new output unit.

Once in a resonant state, ART1 updates its connections by the following:

$$\begin{aligned}\mathbf{w}_{\mathbf{bu}}(\mathbf{i}) &= \frac{L}{L-1+\sum_j \mathbf{F1}'_j}\mathbf{F1}' \\ \mathbf{w}_{\mathbf{td}}{}^{\mathrm{T}}(\mathbf{i}) &= \mathbf{F1}'\end{aligned} \tag{B4}$$

where *L* is a free parameter and *i* is the position of the winning unit.

Expectancy of ART1's activation

The problem of presenting random vectors to an ART1 architecture (Carpenter & Grossberg, 1987) is akin to the well-known urn problem in probability (Ross, 1998). In the urn problem, *n* draws are made in an urn containing *N* balls, *m* of which are white and the remaining (*N* - *m*) are black. Let's define *X* as the number of white balls that are drawn. In this case, *X* is hypergeometric, which is described by:

$$P(X=i)=\frac{\begin{pmatrix} m \\ i \end{pmatrix}\begin{pmatrix} N-m \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{B5}$$

In the present simulations, the cardinality of the stimulus vector is the number of balls in the urn (*N*), the number of draws is equivalent to the number of positions in the random vector containing non-null values (*n*), and the number of white balls is the number of non-null values in the weight vector (*m*). Because Eq. B1 makes the random vectors binary, *n* is binomial with parameters (*N*, 1/2). Therefore, the expectation of the activation of a given unit is described by:

$$E(a)=\frac{L}{L-1+m}\left(\frac{1}{2}\right)^{N}\sum_{n=0}^{N}\begin{pmatrix} N \\ n \end{pmatrix}\sum_{i=0}^{n}\frac{\begin{pmatrix} m \\ i \end{pmatrix}\begin{pmatrix} N-m \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}i \tag{B6}$$

In the second simulation set, *L* = 45 and *N* = 256. Eq. B6 shows that the expectation of the activation of an output unit is a non-linear positive monotonic function of *m*. Since *m* = {59, 50, 32} for category “1”, “0”, and “7” respectively, the expectation of the activation of the “1” category is higher than the expectation of the “0” category, which is higher than the expectation of the “7” category. Because ART1 is a competitive learner, the maximal activation is always chosen, which explain the network’s bias.

Table 1.
Classification of the random vectors

	RAMs			Competitive Networks		
	"A"s	"B"s	χ^2	"A"s	"B"s	χ^2
Uniform	153 (0.46)	178 (0.54)	0.593	237 (0.47)	263 (0.53)	0.512
Bimodal	220 (0.44)	276 (0.56)	2.915	246 (0.49)	254 (0.51)	0.200
Step	195 (0.64)	110 (0.36)	7.053	238 (0.48)	262 (0.52)	15.75*
Exponential	357 (0.72)	142 (0.28)	3.085	221 (0.44)	279 (0.56)	222.4*

Note. 500 random vectors were tested in each cell. Vectors not appearing in the RAMs' conditions stabilized in spurious states. Numbers in parenthesis represent proportions. * indicate results that significantly differed from their training distribution according to a $\chi^2(1)$, $\alpha = .001$. The critical value was 10.83.

Table 2.

Classification of the random vectors

	NDRAM				ART1			
	"7"s	"0"s	"1"s	χ^2	"7"s	"0"s	"1"s	χ^2
Uniform	170 (0.34)	145 (0.29)	185 (0.37)	5.089	6 (0.01)	134 (0.27)	360 (0.72)	375.9*
Multimodal	172 (0.34)	145 (0.29)	183 (0.37)	6.095	5 (0.01)	133 (0.27)	362 (0.72)	382.8*
Step	231 (0.46)	134 (0.27)	135 (0.27)	5.207	8 (0.02)	137 (0.27)	355 (0.71)	653.6*
Exponential	251 (0.50)	143 (0.29)	106 (0.21)	1.533	9 (0.02)	114 (0.23)	377 (0.75)	922.9*

Note. 500 random vectors were tested in each cell. Numbers in parenthesis represent proportions. * indicate results that significantly differed from their training distribution according to a $\chi^2(2)$, $\alpha = .001$. The critical value was 13.82.

Figure captions

Fig. 1. Stimuli used in the first simulation set. All stimuli in the first row (category "A") are orthogonal to all stimuli in the second (category "B").

Fig. 2. (a) Block diagram representing the architecture of the RAM. **W** is the weight matrix representing a simple linear autoassociator, $\mathbf{x}_{[t]}$ is the state of the network at time *t*, and the grey square is a delay unit. (b) Feedforward architecture of the competitive network. The dashed arrow represents inhibitive connections (with no adjustable weights).

Fig. 3. Distributions used as environments in the first set of simulations. The first three stimuli were from category "A" and the remaining from category "B" (see Fig. 1). (a) Condition in which every exemplars and every categories were equally likely. (b) Condition in which all categories were equiprobable but some exemplars were biased. (c) Condition in which the first category was biased but all exemplars in a category were equiprobable. (d) Condition in which the first category was biased, and the probability associated to each stimulus was inversely related to its position.

Fig. 4. Eigenvalues of the weight matrices developed by the RAMs after training. In each panel, the first eigenvalue was associated to "A"s and the second to "B"s. Panels represent the same conditions as in Fig. 3.

Fig. 5. Eigenvectors of the weight matrices developed by the RAMs (after convergence). In each panel, the top eigenvector represent "A"s and the bottom "B"s. Panels represent the same conditions as in Fig. 3.

Fig. 6. Weight matrices developed by the competitive networks. Panels represent the same categories and conditions as in Fig. 5.

Fig. 7. Stimuli used in the second simulation set. The first row shows handwritten "7"s, the second row handwritten "0"s, and the third row "1"s. The stimuli were coded as 256-units bipolar

vectors (±1). For reference, stimuli were numbered from left to right and from top to bottom.

Fig. 8. (a) Architecture of NDRAM. (b) Architecture of ART1. The circles represent gating mechanisms and the rectangles layers of neurons. The left part of ART1 is the network per se and the right part is the orientating subsystem (novelty detector).

Fig. 9. Distributions used as environments in the second set of simulations. The first ten stimuli were “7”s, the following ten were “0”s, and the remaining were “1”s. The panels have the same properties as in Fig. 3.

Fig. 10. First twenty eigenvalues of the weight matrices developed by NDRAM after training. Panels represent the same conditions as in Fig. 9.

Fig. 11. Attractors developed by NDRAM. Panels represent the same conditions as in Fig. 9.

Fig. 12. Weight matrices developed by ART1 ($\mathbf{W_{bu}}$). In each panel, the leftmost weight vector was associated to the first output unit, the middle weight vector was associated to the second unit, and the rightmost to the third output unit. Panels represent the same conditions as in Fig. 9.

Figure 1

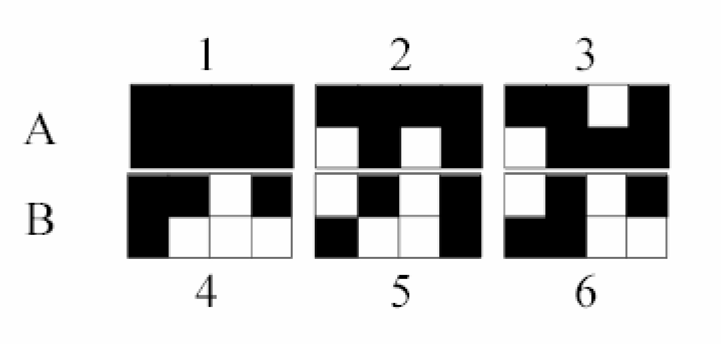
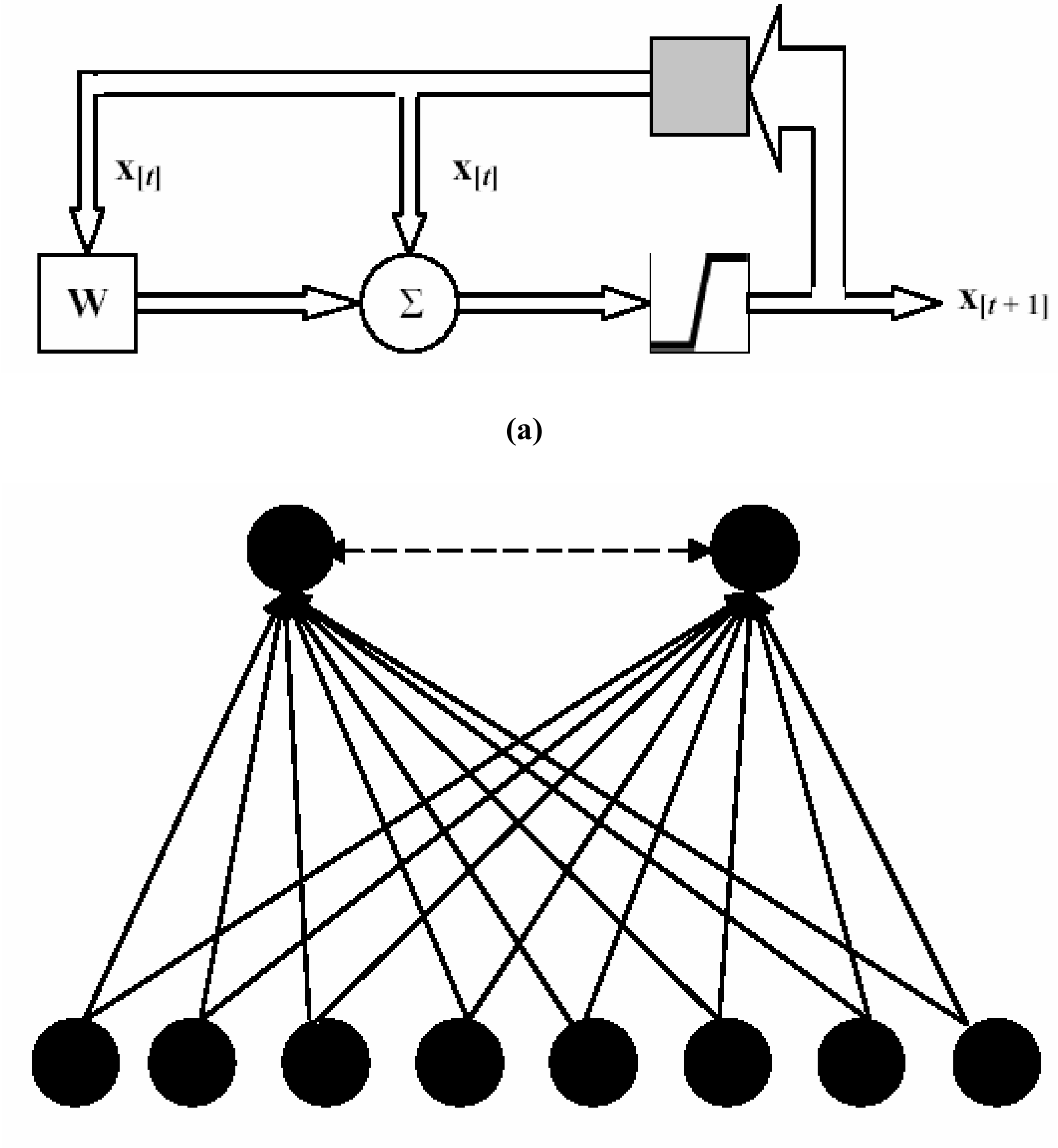


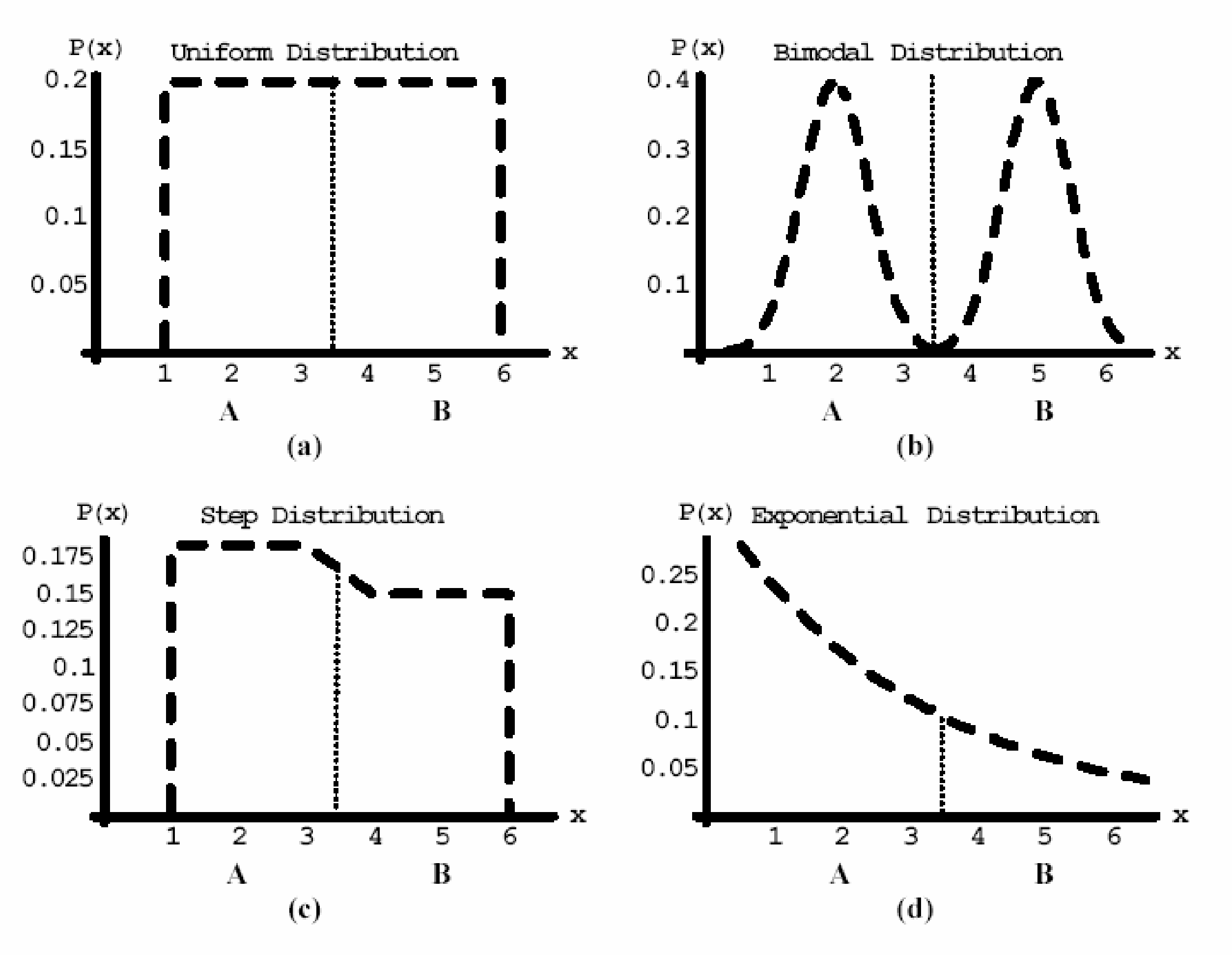
Figure 2



(a)

(b)

Figure 3



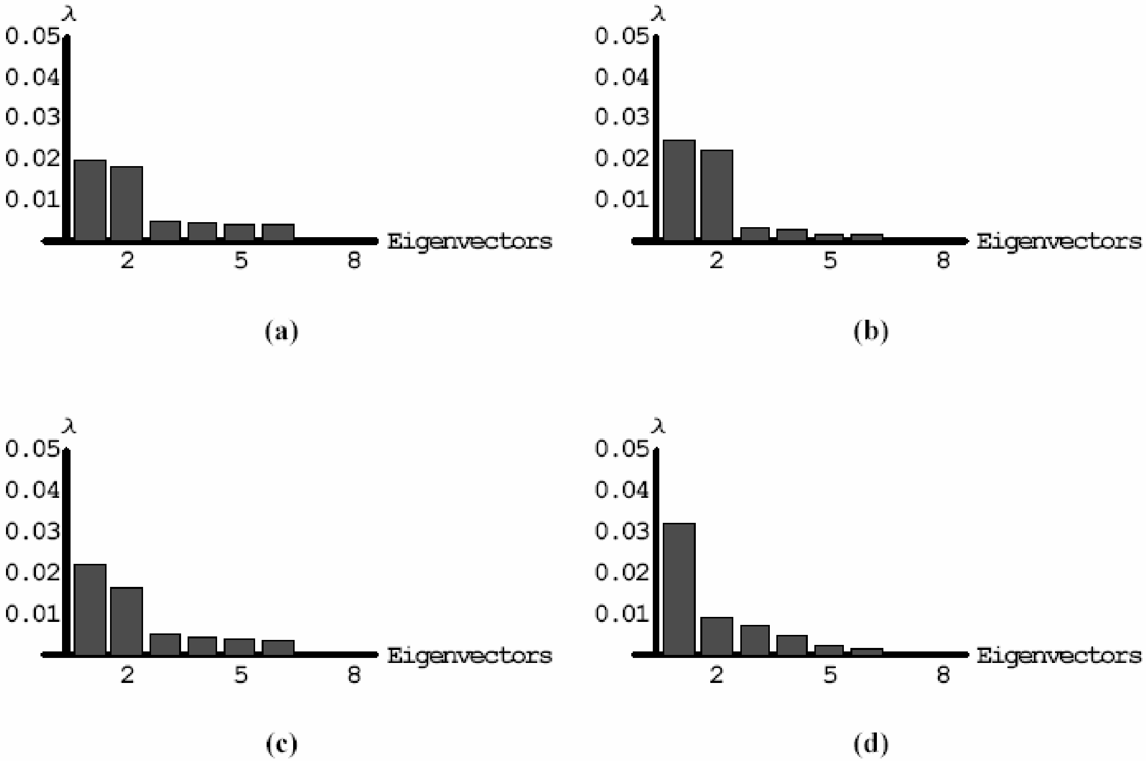
(a)

(b)

(c)

(d)

Figure 4



(a)

(b)

(c)

(d)

Figure 5

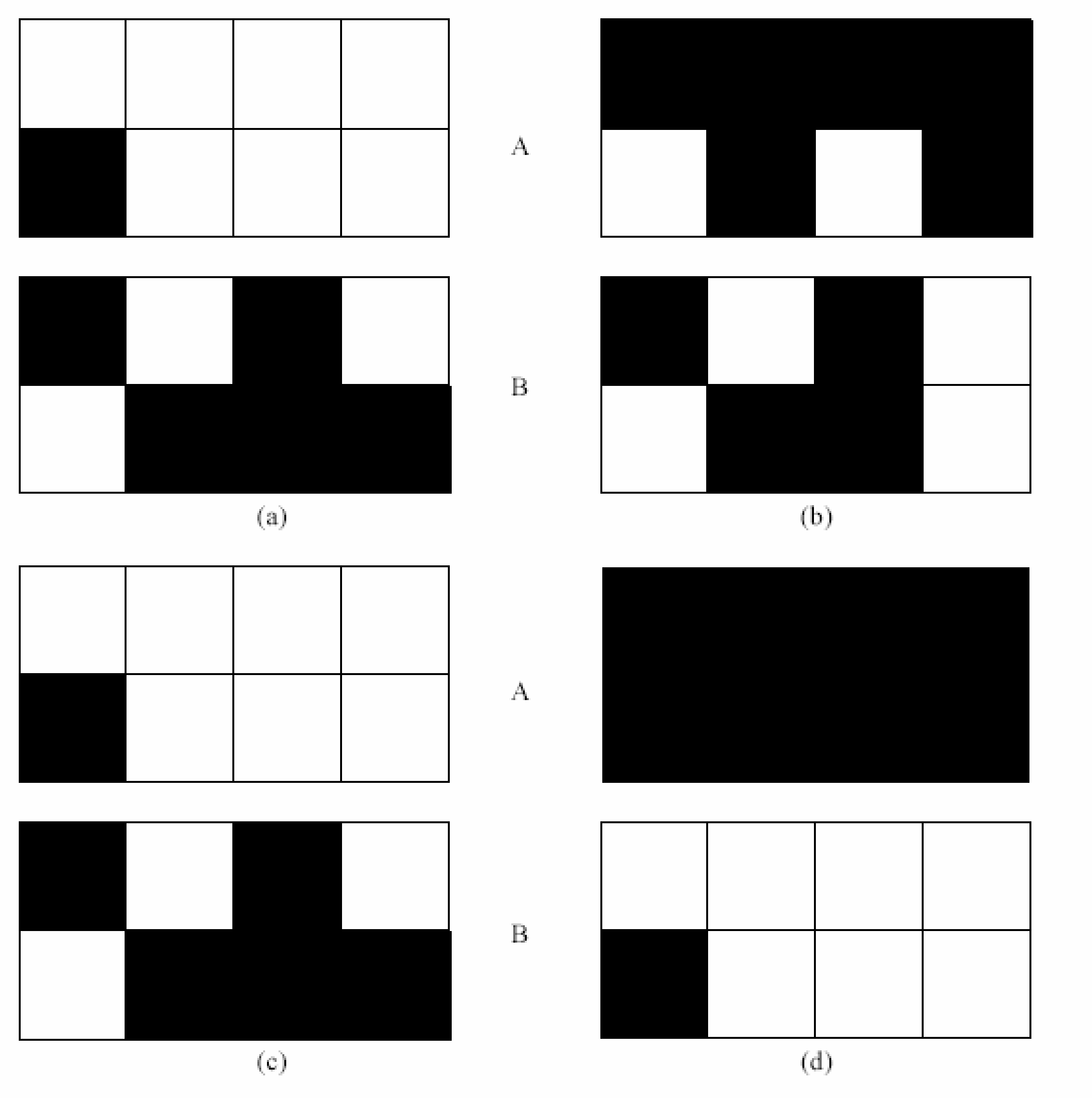
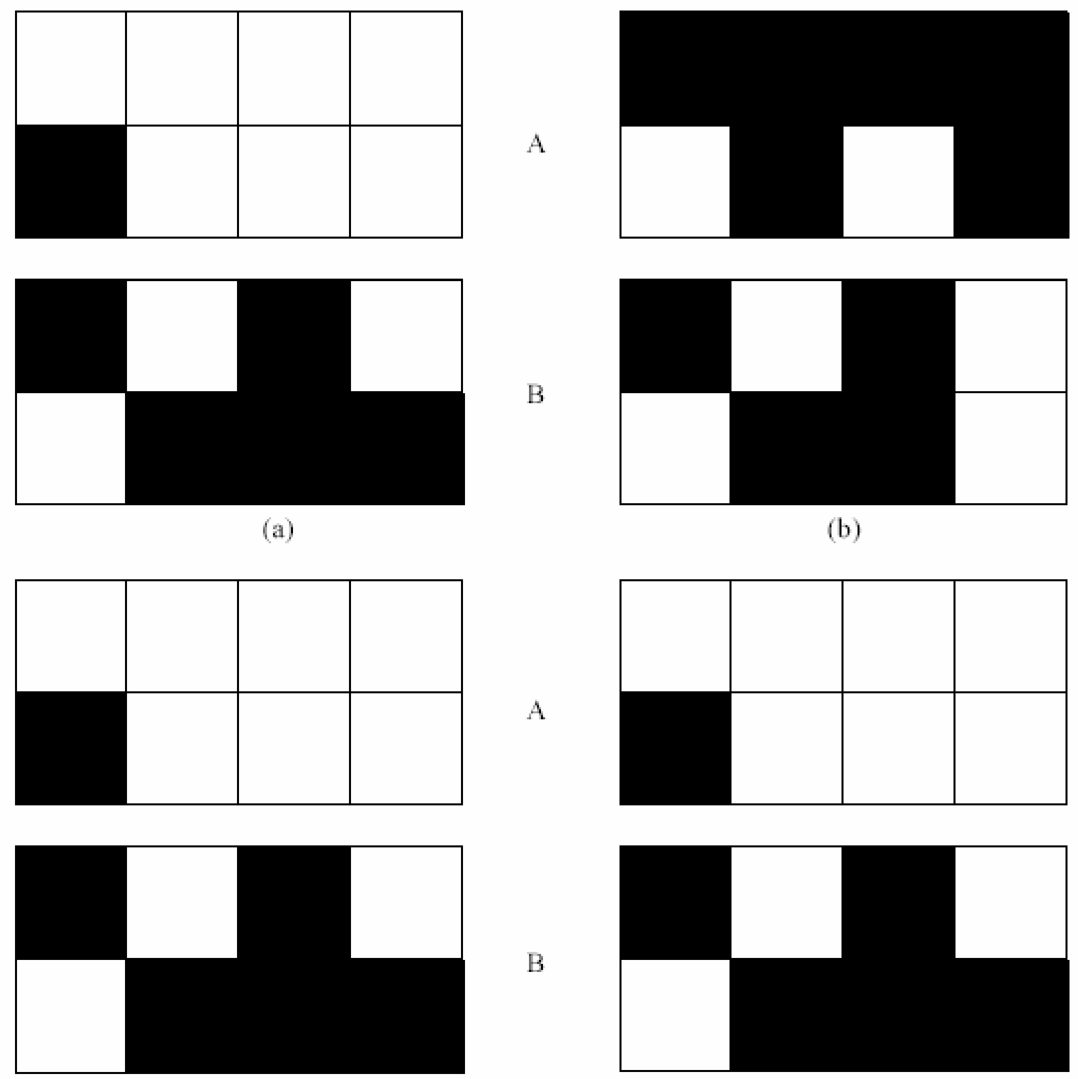


Figure 6



(a)
(b)
(c)
(d)

Figure 7

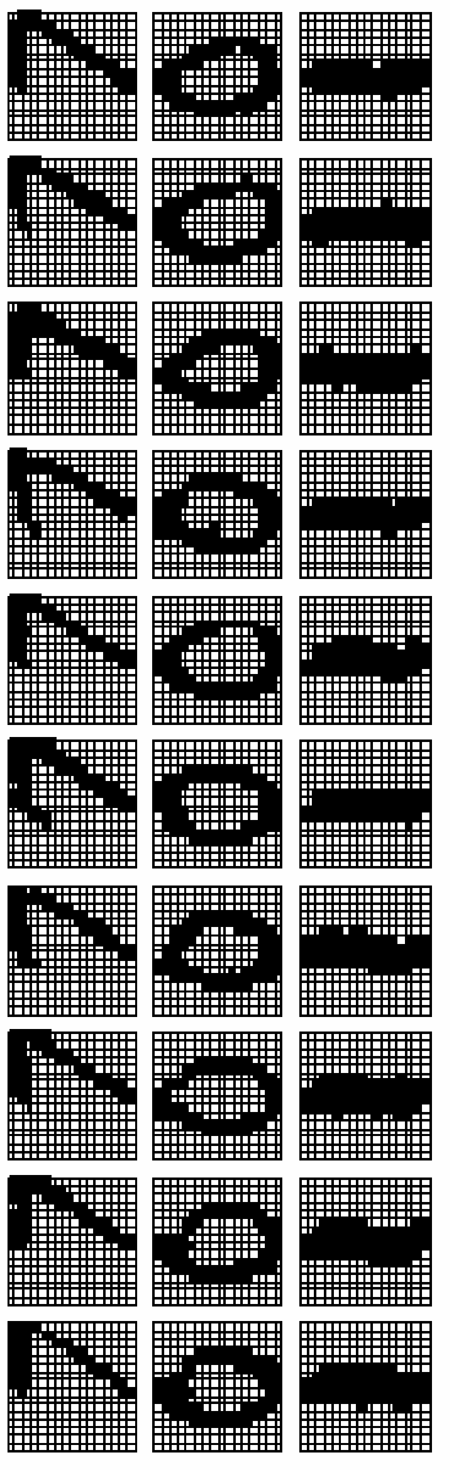
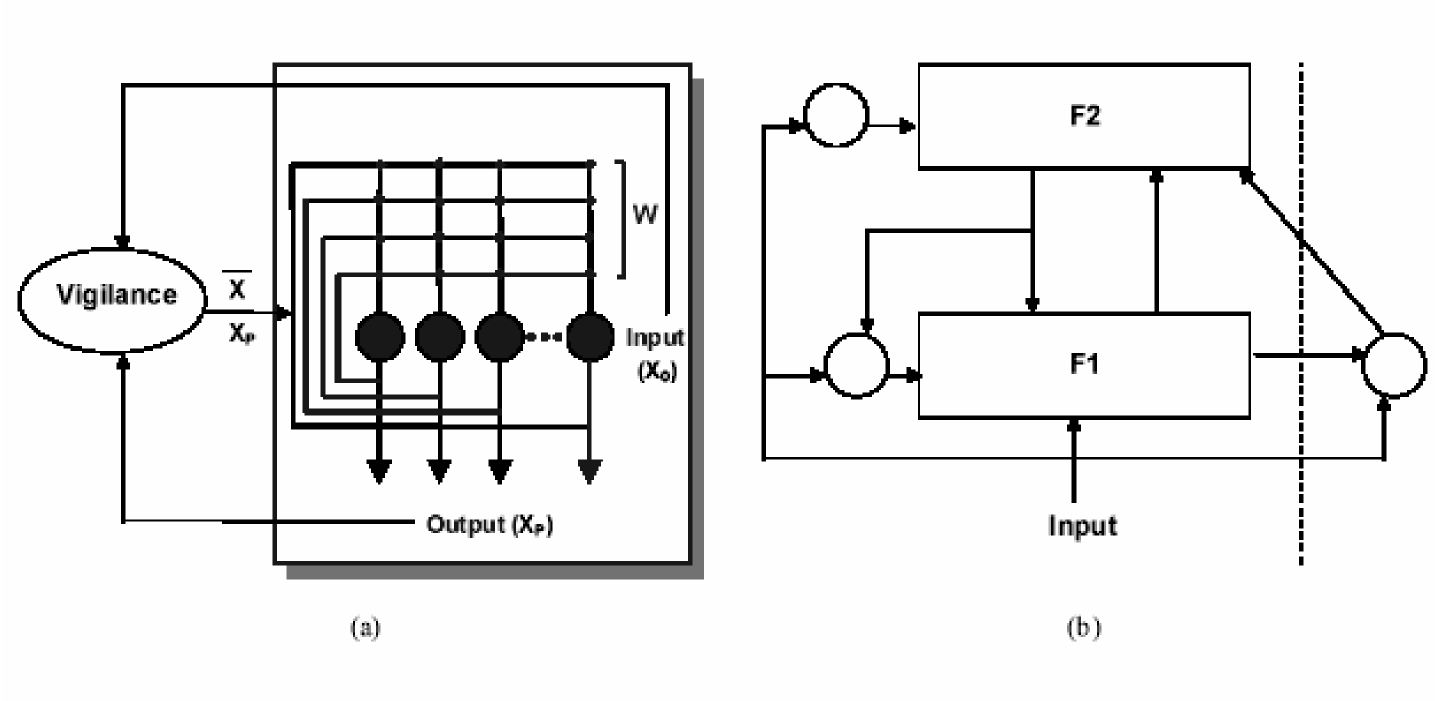


Figure 8



(a)

(b)

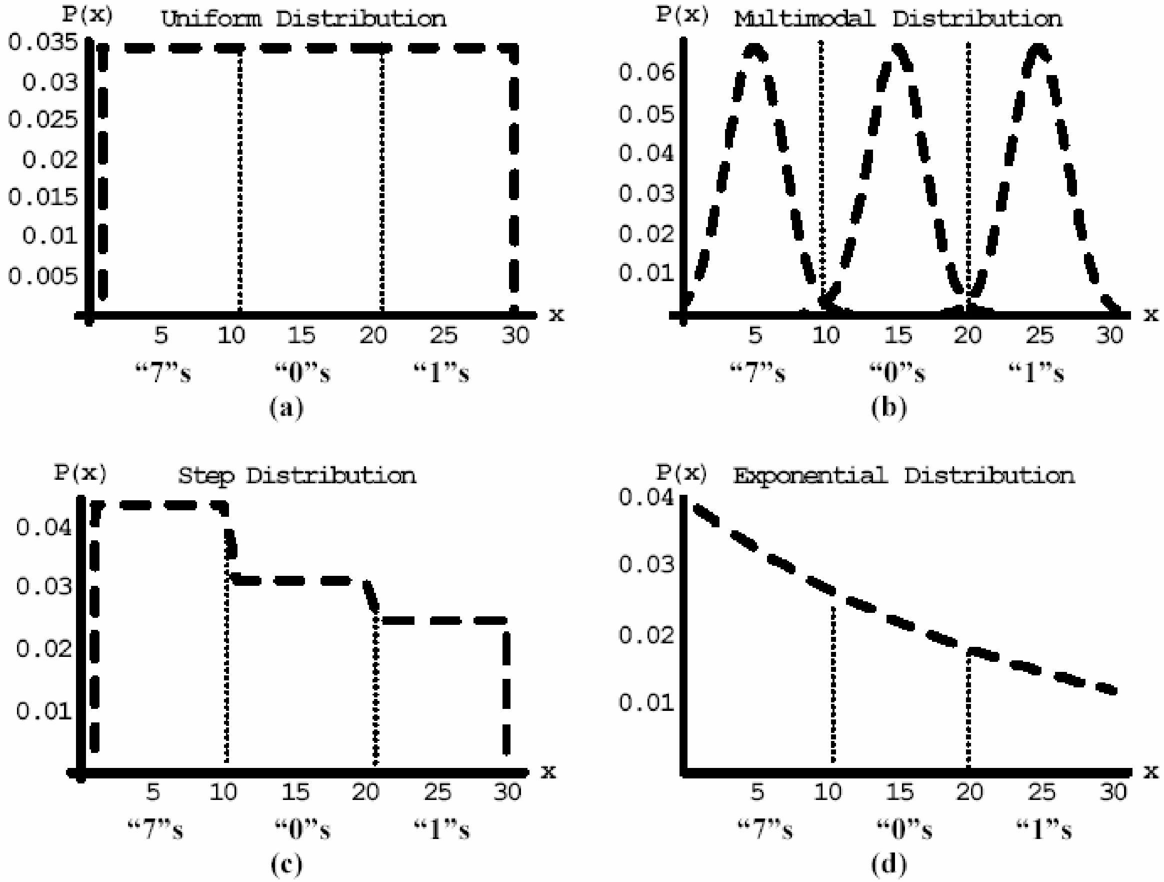
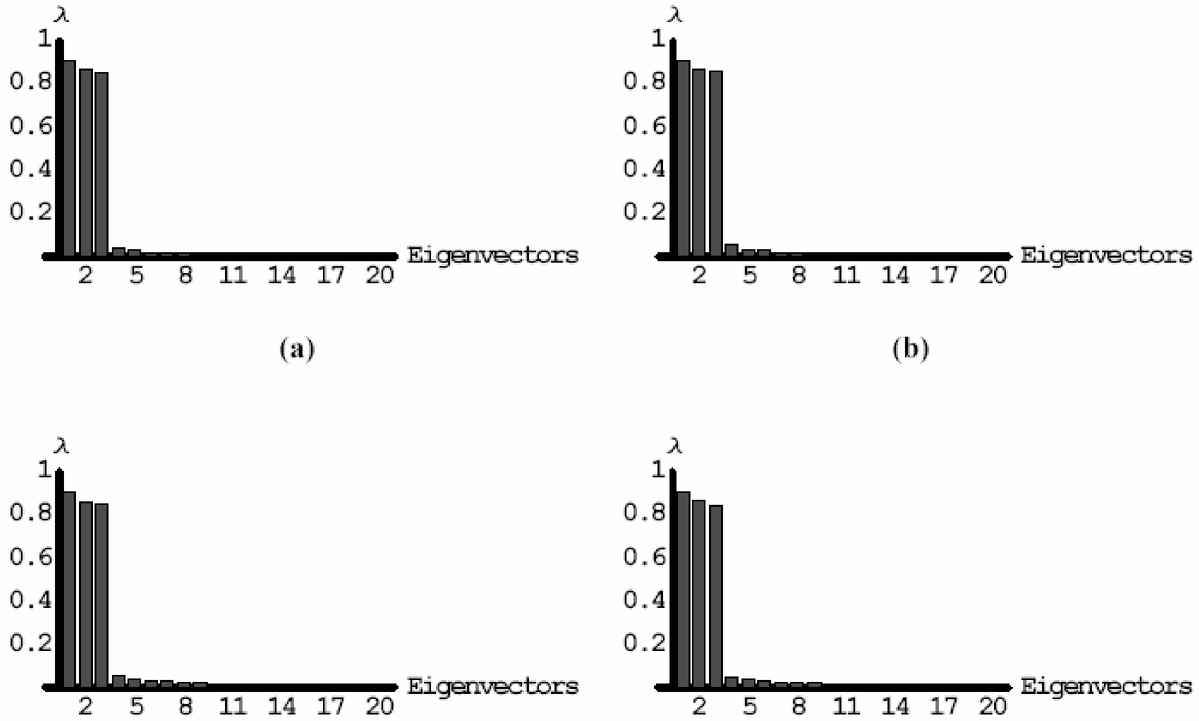

P(x)
Uniform Distribution
0.035
0.03
0.025
0.02
0.015
0.01
0.005
5
10
15
20
25
30
x
"7"s
"0"s
"1"s
(a)
P(x)
Multimodal Distribution
0.06
0.05
0.04
0.03
0.02
0.01
5
10
15
20
25
30
x
"7"s
"0"s
"1"s
(b)
P(x)
Step Distribution
0.04
0.03
0.02
0.01
5
10
15
20
25
30
x
"7"s
"0"s
"1"s
(c)
P(x)
Exponential Distribution
0.04
0.03
0.02
0.01
5
10
15
20
25
30
x
"7"s
"0"s
"1"s
(d)

Figure 10



(a)

(b)

(c)

(d)

Figure 11

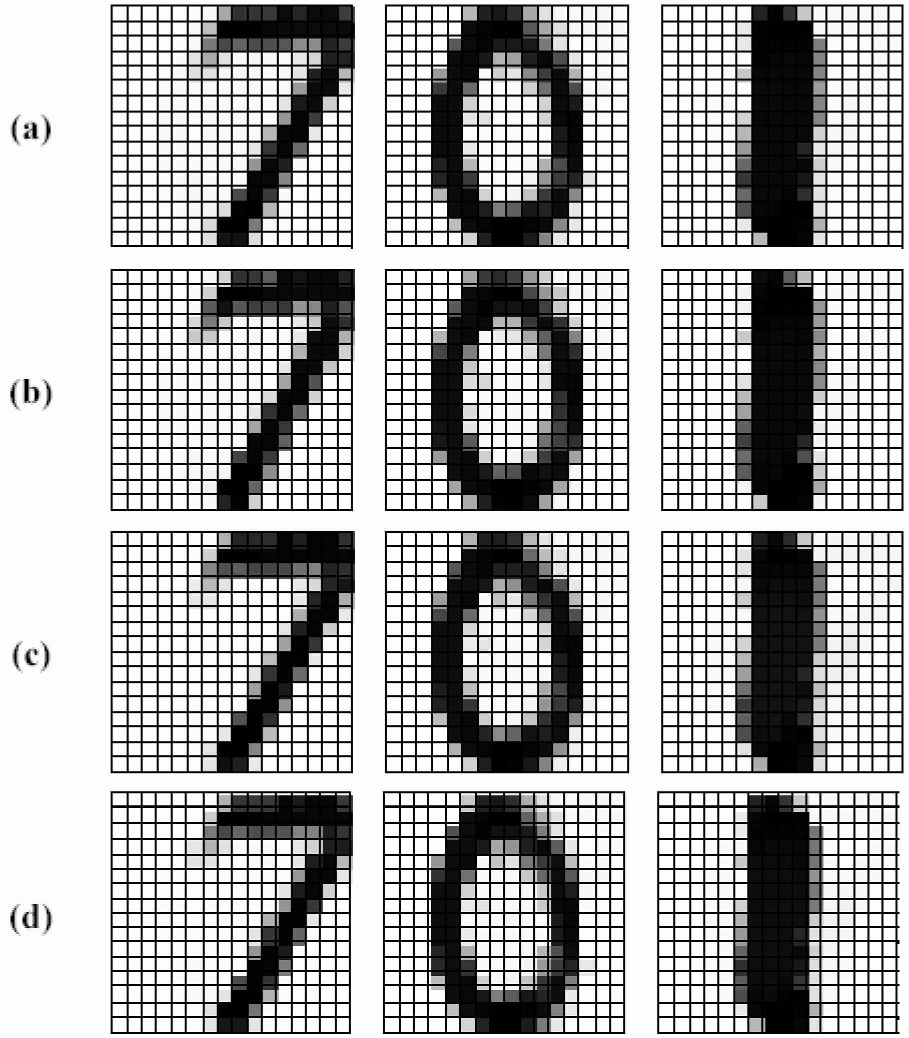


Figure 12