A Neurodynamical Model of Context-Dependent Category Learning

Laxmi R. Iyer and Ali A. Minai, *Senior Member, IEEE*

***Abstract*—The abstraction of patterns from data and the formation of categories is a hallmark of human cognitive ability. As such, it has been studied from many different perspectives by researchers, and these studies have led to several explanatory models. In this paper, we consider the inference of categorical representations for the purpose of producing task-specific responses. Task-relevant responses require a knowledge repertoire that is organized to allow efficient access to useful information. We present a neurodynamical system that infers functionally coherent categories from semantic inputs (or concepts) presented sequentially in different contexts, and encodes them as attractors in a two-dimensional topological feature space. The resulting category representations can then act as pointers in a larger system for semantic cognition. The system allows controlled hierarchical organization and functional segregation of the inferred categories.**

I. Introduction

The brain receives a vast amount of sensory information from a variety of sources. Creating patterns and structures out of this information and using these patterns to make further inferences and predictions is a key feature of cognition. The mechanisms underlying this important ability are still not fully understood, and doing so is a major challenge not only for cognitive scientists seeking to understand brain function but also for computer scientists trying to build artifical intelligent systems. The performance of computers on higher order cognitive functions still does not approach that of even simple animals, and better neurocomputational models of categorization will help greatly in overcoming this limitation.

Category learning has been studied extensively by psychologists and neuroscientists, which has led to the realization that such learning involves several distinct systems [3].

Early theories of categorization focused on *rule-based category learning*, where categories represent simple conjunctions of features that can be described explicitly as a rule or strategy. This type of category learning has been associated with the prefrontal cortex (PFC) and the anterior cingulate cortex (ACC), which have also been implicated in executive function and working memory [32]. It has become clear, however, that rule-based learning cannot account for all types of categorization. This has led to theories of *prototype-based categorization* [28] and *exemplar-based categorization* [19], [22]. Prototype-based categorization defines a category in terms of typical prototype around which specific instances of the category are distributed. Such categorization is associated with regions of the occipital lobe [1], [29], [33]. Humans are also able to handle exceptions to rules and explicit representations of category members. Such categorization is termed exemplar based categorization, and has been associated with the medial temporal lobe (MTL) [3], [32]. Some studies have shown that there is a competition between the PFC and the MTL during categorization tasks [32].

Computational models that have studied category learning have modeled all three types of categorization, as well as several hybrids. Rule-based categories have been modeled by RULEX [24] and COVIS [2]. Most computational models – including popular models like ALCOVE [13] and SUSTAIN [17] – model exemplar and prototype based categorization. The computational model of category learning dsescribed in this paper is closest to rule-based category learning. This model is embedded in a larger framework of executive function, which has been described elsewhere [11], [10].

Outside laboratory tasks, humans seldom learn categories simply for the sake of doing so. Categories are learned mainly in order to facilitate the performance of other functions [18] such as ideation, inference, choice or action selection, etc. [32]. We have recently proposed a model where categories act as pointers in the selection and activation of context-appropriate functional repertoires [11], [10], [12]. According to this model, cognitive functions such as ideation, motor control, recall, etc. can be seen conceptually in terms of four core systems:

- **Long-term memory (LTM) systems** that encode information and methods relevant to cognitive function in general, independent of the context.
- **An internal attention and working memory (IAWM) system** that focuses attention on context relevant knowledge and methods.
- **A reward system** that generates reward based on task-appropriate responses.
- **Modulatory systems** that modulate the activity of the IAWM and LTM systems based on reward.

In the model, the LTM comprises a set of LTM-items, organized into LTM modules representing *categories*. These group together LTM-items sharing features that are salient to the task, i.e. LTM-items that are similar *for the purposes of that task or context*. A *response* comprises a combination of

Laxmi Iyer is with the Department of Computer Science, University of Cincinnati, Cincinnati OH 45221, Email: iyerlr@email.uc.edu. Ali Minai is with the Department of Electrical and Computer Engineering, University of Cincinnati, Cincinnati OH 45221, Email: Ali.Minai@uc.edu

Acknowledgement: This work was supported in part by a National Science Foundation Human and Social Dynamics Program grant to Ali Minai (BCS-0728413), which includes support from the Deputy Director of National Intelligence, and by a National Science Foundation CreativeIT grant to Ali Minai (IIS-0855714). The views expressed in this paper are not those of the NSF.

LTM-items. It is assumed that there is a tonic inhibition of the LTM categories by default [25], [32], and categories must be disinhibited to become available for generating responses. The categories that are selectively disinhibited by the context bias their associated LTM-items. This forms a context-dependent knowledge repertoire within which appropriate responses can emerge. The LTM categories thus play an important role in ensuring the efficient and effective generation of context-appropriate responses. The model described in this paper focuses on the learning of these LTM categories.

The next section provides a brief review of the prominent theories and models of category learning. This is followed by a description of our model and its motivation. The next section gives a description of our methodology and implementation of the system, followed by system simulations and results.

II. Background

As discussed earlier, theories of category learning can be divided broadly into two types: 1) Rule-based learning, which uses rules or strategies to define categories; and 2) Similarity-based learning, which defines membership based on similarity of items to each other or to a prototype. Similarity-based schemes can be sub-divided further into prototype-based and exemplar-based.

Both neuropsychological patient data and neuroimaging data show that rule based category learning is associated with the prefrontal cortex (PFC) and anterior cingulate – structures that are associated with executive function and working memory. Miller and colleagues and others have found that cells in the PFC can encode rules, responding to very different stimuli that satisfy the rule, but not to otherwise similar stimuli that do not satisfy it [20], [6].

In prototype-based categorization, a category is described in terms of a central tendency, i.e., the tendency of exemplars to cluster around a prototype. This type of categorization has been demonstrated by many experimental studies e.g. [28], [8], [30], [9], [34], [35]. In several of these, subjects classified the prototype as a member of a category more frequently than they did new members, or even existing members e.g.[28], [8]. Neuroimaging data show that prototype based-category learning is associated with changes in the occipital cortex [1], [29], [33].

Exemplar-based categorization is a variant of prototype-based theories where categories are defined based on the similarity of stimuli to specific exemplars of each category [5]. These exemplars need not be global prototypes in the sense of defining a cluster centroid, but rather serve as local prototypes. This type of categorization can handle categories with complex boundaries in representational feature space. It is associated with the medial temporal lobe (MTL) [3], [32].

A. Computational Models of Category Learning

Several computational models of category learning have been developed, both by cognitive scientists and by artificial intelligence researchers. These are briefly reviewed in this section.

A family of computational models is based on the generalized context model (GCM) [21], which is a formal generalization of the context model of Medin and Schaffer [19]. Given a new stimulus, the model matches it with an existing stimulus using a Bayesian rule, which estimates the probability of the category given the stimulus. Extensions of GCM include the exemplar based random walk (EBRW) model [23], where exemplars direct a random walk process that leads to the categorization of the stimulus. In the extended generalized context model (EGCM), only salient dimensions of exemplars are considered during categorization [14], [15].

An important extension of the generalized context model is the ALCOVE model [13]. ALCOVE is a connectionist implementation of the GCM, where the attentional weights of each dimension are learned adaptively. ALCOVE is implemented as a three-layer feedforward neural network, with the first layer consisting of stimulus features, the second or hidden layer consisting of individual exemplars, and the output layer encoding the categories. Weights between the stimulus and hidden layers are attentional weights, which are adaptively learned, while weights between the hidden and output layers are association weights. A model of category learning similar to ALCOVE is the SUSTAIN model [17]. This is a clustering model and its basic principle is that the category representation formed by humans depend strongly on the tasks and goals of the learner. The model is initially directed towards simple solutions, but moves towards more complex solutions when these do not work.

Another connectionist feedforward model of categorization is the configural cue model [26], [27], which is implemented as a two-layer feedforward network. The input representation consists of all possible combinations of features, i.e. the power set. The presentation of a stimulus corresponds to the presentation of all the possible subsets of that stimulus. The activations of the input nodes are multiplied by the connection weights to form the activations of the outputs. Feedback is used to modify the association weights between the layers.

A neuropsychological theory and formalism of multiple category learning called COVIS has been proposed by Ashby [2]. COVIS assumes that category learning is a competition between two separate systems - verbal and implicit. The anterior cingulate and the PFC are critical in the verbal system, while the implicit system makes use of the caudate nucleus of the basal ganglia. The verbal system learns boundaries of the stimulus dimensions while the implicit system is a perceptron-like decision-based model of category learning that learns to map stimuli to responses.

The system we describe in this paper infers categories based on whether stimuli with shared features are relevant within the same context. Thus, it falls broadly within the scope of rule-based category formation systems. It is also seen as part of a larger system where the inferred categories are used to construct context-specific response repertoires, which is a function associated with cognitive control and working memory. As such, the model shares some features

with the SUSTAIN model. The model is described in greater detail next.

III. Features of the Model

Categories are most useful when they do not merely capture the statistical structure of the data, but are based on goal and task requirements [17]. Several studies show that during categorization, people form representations that are useful for a specific task. If given extensive practice with a new task, they will form a distinct categorical representation for it[18]. However, studies also show that representations formed in one task affect the learning of subsequent tasks, showing that categories formed for specific tasks are then applied more generally. In our model, we postulate that the learning of categories be *task-specific* while its usage is *task independent*. The basic principle defining categories in our model is that each category must group together items or concepts that are functionally substitutable and occur frequently enough in one or more specific contexts. The requirements for the resulting model are as follows:

1. **The extraction of relevant subspaces for categorization**: On being trained with many exemplars, the system should be able to recognize a subset of frequently co-occuring features and learn it as a category. This form of abstraction of salient features from many exemplars has been noted in several psychological studies, e.g. [7].
2. **Context-dependent learning**: The system must be able to pick out only those combinations of frequently co-occuring features that are disproportionately prevalent in specific contexts. The role of the model is not to pick out the statistical regularities in the data in general, but to pick out the regularities that are useful in specific tasks or contexts.
3. **Context-independent usage**: Although the system learns only categories that are useful in some contexts, the system should be able to use these as a generic pool of categories that are useful across all contexts.
4. **Functional gating**: Since the objective is to form categories that are considered equivalent for some task, the concepts in a category should all serve a common function, i.e., they should group together concepts that serve the same function.

IV. Problem Formulation

The LTM comprises a set of N_c *concepts*, $C = [C^1, C^2, ..., C^{N_c}]$, each represented in terms of its *features*. The system has N_f features which are divided into two sets: a) A set, F^T, of N_f^T *type features*, which indicate the type or function of concepts, such as 'food', 'tool', 'animal', 'activity', etc.; and b) A set, F^D, of N_f^D *descriptive features* which represent attributes of concepts. The complete set of features is denoted as $F = F^T \bigcup F^D$. Distinct features represent opposing or mutually exclusive attributes [36]. The units are binary. Each concept has only one active type feature but can have several attribute features with non-zero values. Thus, the feature representation for a concept i is the vector $\Phi_i = \{\Phi_i^T \ \Phi_i^D\}$, where $\Phi_i^T = \{\phi_{ij}^T\}$ is the type feature vector with one $\phi_{ij}^T = 1$ and the rest 0, and $\Phi_i^D = \{\phi_{ij}^D\}$ is the attribute feature vector.

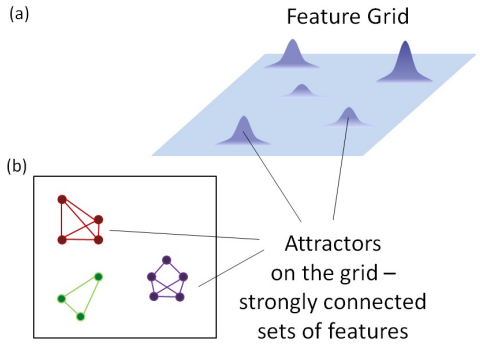

Fig. 1. Slow learning - As the system is trained on concepts, the feature grid slowly develops localized attractors (See (a)), each of which consists of a strongly connected set of features. (See (b))

The system experiences several task *contexts*, each exposing it to a sequence of concepts. Each concept is presented to the system individually as a feature vector. The goal for the LTM is to infer categories by grouping together concepts that: 1) Are of the same type; and 2) Share a subset of features that co-occur repeatedly within the same context. Formally, a category s_k is defined by a set of *defining features* [37] comprised of one *type feature*, f_t^k and several *descriptive features* $[f_{d_1}^k, f_{d_2}^k ... f_{d_{N_k}}^k]$, such that all concepts in the category have the type feature $f_t^k = 1$, and the descriptive features $[f_{d_1}^k, f_{d_2}^k ... f_{d_{N_k}}^k]$ all have values of 1.

A subset of type-attribute features that co-occur often within one or more contexts is said to be *salient*. The system's task is to detect such salient feature combinations and to construct a robust category representation for each such combination. The categories obtained as a result have the following properties:

1) Pairs of concepts sharing a set of features that are salient in one or more contexts are grouped into a single category even though they may differ in many other features.
2) Pairs of concepts that do not share features that are salient in any context are assigned to different categories even though they may share a large fraction of their features.
3) Concepts in different contexts that share the same set of salient features are grouped into the same category.
4) Concepts in different contexts that share similar but not identical sets of salient features are assigned to distinct categories.

Very importantly, the learned categories are encoded as local attractors (similar to bump attractors) so that they can be activated correctly by noisy versions of concepts after learning.

V. System Description

The system comprises two 2-dimensional layers of neural units: A *feature layer*, F, which encodes category represen-

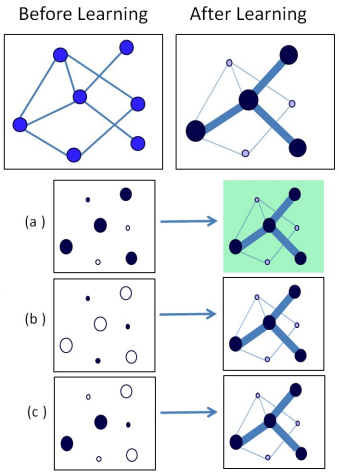

Fig. 2. After learning, weights between prominently occuring features is relatively high, while weights between other features is relatively low compared to the unlearned weights. Also each frequently activated unit becomes habituated to the amount of input it gets. As a result, when a concept that contains the correct set of features is presented, the system settles to the attractor (See (a)) The units in the big circles are the inputs. The dark circles denote the salient features, while the hollow circles denote other features.). When the concept that does not contain the correct set of features is presented, the attractor is unlikely to get activated (See (b)). When a concept contains part of the correct feature units, the attractor is unlikely to get activated (See (c)). So the system acts as an attractor to the correct input and a repeller to the wrong input and to partial input.

tations, and a *modulation layer*, M, which modulates the synaptic plasticity in the feature layer.

The feature layer is a 2-dimensional grid of *feature units* receiving the current concept vector as input. Each unit, i, is tuned to a specific feature, $\mu(i) \in F$, with multiple units for every feature distributed randomly across the grid. The pattern of tuning across the grid is such that units for all features can be found in every local neighborhood of the grid, albeit in random relative positions. The feature units are connected recurrently to one another with local excitatory connections and global inhibitory connections. The presentation of a concept to the system activates all units tuned to features present in the concept. Each activated feature unit excites units within its local region while inhibiting other units. This results in competitive activity leading to the concentration of activity in a local region where the units corresponding to the concept's features are especially close to each other. These localized activity patterns are reinforced through a variant of competitive Hebbian learning where the weights between co-active local units are increased strongly while the remaining local connections are strongly depressed. The modulation from the integration layer ensures that this depression is confined to regions with strong activity. This results in the formation of strong localized attractors, each activating a set of units tuned to frequently co-occuring features (See Figure 1). These attractors encode categories.

In addition to the strongly divergent synaptic modification, units that are active in a learned attractor also increase their activation thresholds, i.e., become harder to activate. These two effects combine to ensure that, as they are learned, attractors become increasingly tuned to specific feature combinations and cannot be activated by concepts that match them partially. Thus, the presentation during learning of a concept that matches already learned attractors weakly leads to the formation of a new attractor in another region, while the presentation of a concept strongly similar to an existing attractor simply reinforces that attractor further. Over time, this leads to the emergence of distinct attractors for categories that differ in at least some salient features. (See Figure 2). This is important because it enables the system to form several localized attractors in different regions of the grid.

Another feature of the system is that, until weights reach a *stabilization threshold*, υ, or fall to 0, they revert gradually towards their baseline values in the absence of potentiation or depression. Thus, stable attractors can only be formed for salient feature combinations that co-occur with sufficient frequency in time, i.e., within the same context (or rarely in two neighboring ones). This ensures that each category learned by the system has sufficient frequency within at least one context. Feature combinations that occur often but are thinly spread across many contexts – and thus widely separated in time – are *not* learned as categories. The categories learned by the system over time are, therefore, context-relevant.

VI. System Implementation

A. Initialization of the Feature Layer

The 2D feature grid consists of N_F feature units and is assumed to have periodic boundary conditions to avoid edge effects. Features are assigned to the units in the grid so as to maximize the distance between units representing the same feature, and to ensure that each neighborhood on the grid has a wide diversity of features. Such a locally diverse distribution of feature detectors was proposed by Barlow as a fundamental principle of neural organization [4].

Each feature unit i has two dimensional coordinates $\mathbf{q}_i = (x_i,\ y_i)$ on the grid. The baseline weights between units i and j depend on the Euclidean distance, d_{ij}, between them, and have a modified Mexican hat form. They are calculated as follows. A scaled distance $\bar{d}_{ij} = d_{ij}/r$, is calculated, where r is a system parameter that controls the width of the excitatory field around i. Units that have $d_{ij} < r$ (i.e., $\bar{d}_{ij} < 1$) have excitatory connections, while those further apart have global inhibitory connections.

The excitatory weight from unit j to unit i is calculated as:

$$W_{ij}^E = \begin{cases} \frac{2(1-\bar{d}_{ij}^2)}{\sqrt{3}\pi^{1/4}} exp(-\frac{1}{2}\bar{d}_{ij}^2), & \bar{d}_{ij} < 1; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

and the inhibitory weight as $W_{ij}^I = -1$, if $\bar{d}_{ij} > 1$ and 0 otherwise. The weights between units i and j are thus initially symmetric. The set of units sending excitatory connections to a unit i are termed its *excitatory receptive field*, which has a radius r.

B. Feature Layer Dynamics

The system experiences a stochastic sequence of contexts modeled as a Markov chain, so that each context persists for an exponentially distributed duration before the system transitions to the next context. Concepts are presented sequentially during each context. Each concept, X^k, is represented as a binary vector $[x_1^k, x_2^k, ..., x_n^k]$, where $x_j^k = 1$ if the concept has feature f_j and 0 otherwise. When applied to this grid, the concept excites all feature units tuned to features active in the given concept, creating an *initial mask pattern*, $C^k = [c_1^k,\ c_2^k,\ ...\ c_{N_F}^k]$ where $c_i^k = 1$ if $x_j^k = 1$ and $\mu(i) = f_j$; otherwise, it is 0.

Within each context, the system operates on two time scales: The faster time-scale, indexed by t, represents the cycles of the recurrent network, whereas the slower one, indexed by T, corresponds to the presentation of each new concept to the network. The number of t-steps in each T-step is denoted by n_T and represents the natural "cognitive cycle time" of the system. We term this a *C-cycle* (concept cycle, or cognitive cycle). In a broad sense, T may be seen as corresponding to the theta rhythm and t to the alpha rhythm in the cortical-hippocampal system. Based on experimental data, it has been proposed that the theta rhythm may correspond to the updating of the brain's cognitive state (e.g., by sampling a new stimulus), while the gamma rhythm represents the coordinated dynamics of recurrent cortical activity [16], [31]. The theta rhythm is known to modulate the activity – and, presumably, the excitability and internal state – of neurons. In the model, we assume that the global inhibition on the feature units is modulated by the slow cycle, being 0 when a concept is first presented and increasing gradually over the course of the cycle. The effective weight from unit j to unit i at time t is given by:

$$W_{ij}(t) = W_{ij}^E + (1 - \lambda(\tau(t)))W_{ij}^I \tag{2}$$

where $\tau(t) = tmodn_T + 1$ indicates the index of the fast time-step within the current C-cycle. The λ function is defined as $\lambda(\tau+1) = 1-\tau/(n_T+1)$, so it starts at 1 at the beginning of each C-cycle and ends at 0.

A new concept is presented at the beginning of each C-cycle, activating feature units across the whole system. These units interact with each other through recurrent weights over the following n_T fast cycles, leading to the concentration of activity in a small region of the feature layer. At the end of the C-cycle, the weights between units are potentiated or depressed, and the activation threshold of each unit is adjusted based on its average activity. This is followed by the presentation of the next concept.

Two key features of the model are that the activity of feature units is required to be cooperative and stimulus-gated, i.e., only units receiving recurrent input from several other units and tuned to an active feature in the current concept can become active. This functionality is built into the following expression for the net input to feature unit i at time t:

$$s_i(t) = \frac{\beta c_i \sum_j W_{ij} y_j(t-1)}{1 + e^{-(b_i - \alpha)}} \tag{3}$$

where $y_j(t)$ is the output of feature unit j at time t. Variable b_i equals the number of active feature units in the excitatory receptive field of i, and α is a fixed threshold that determines how many active inputs are needed to activate unit i. Parameter β determines the sharpness of this constraint.

In the first time step, K highest units are chosen to be active, K being much larger than the number of units active at the end. The output of feature unit i at time t is:

$$y_i(t) = \begin{cases} s_i, & s_i > \theta_i^s(t) \text{ and } t \leq 5; \\ 1, & s_i > \theta_i^s(t) \text{ and } t > 5; \\ 0, & otherwise. \end{cases} \tag{4}$$

where $\theta_i^s(t)$ is a slowly time-varying firing threshold that is set initially to 0 but rises gradually as the unit becomes habituated to higher activity. As the equation indicates, the units have more graded outputs at the beginning of each C-cycle, but become increasingly all-or-none as the cycle proceeds. The modulation of inhibition and response during each C-cycle together ensure that the intially network-wide activity generated by each context is forced to converge either to an existing attractor (i.e., category) or to a new localized activity pattern that can be learned as a fresh attractor.

C. Function of the Modulation Layer

The role of the modulation layer is to detect the feature layer region that wins the competition for a given concept, and to modulate the synaptic modification between feature units in that region. In particular, it ensures that, as synaptic weights between co-active feature units rise, the remaining weights in the region are depressed disproportionately. Confining this depression to just the winning region ensures that other regions of the feature layer retain the information they have learned as well as their capacity to respond to future inputs.

The modulation layer has the same number of units as the feature layer, and its units – termed *M-units* – are arranged in a 2D grid of the same dimensions. Each M-unit i receives afferent connections from all feature units j such that $d_{ij} < f_{in}$, where f_{in} is called the *fan-in* of unit i. In turn, each M-unit, i, sends modulatory signals to all feature units, j, within a radius f_{out} of itself, where f_{out} is terms the unit's *fan-out*. Outputs from M-units do not affect the activity of feature units directly, but modulate their synaptic dynamics as described below.

The equations for the M-units are as follows:

The incoming weights into M-unit i from feature units j are:

$$W_{ij}^{MF} = \begin{cases} 1, & d_{ij} < f_{in}; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The connection from M-unit j to M-unit i is given by:

$$W_{ij}^{M} = \begin{cases} 1, & d_{ij} < f_{in}; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The input coming into each M-unit j is:

$$s_i^M(t) = \sum_j W_{ji}^{MF} y_i(t-1) \tag{7}$$

$$y_i^M(t) = \begin{cases} 1, & s_i^M(t) - max_j((s_j^M(t) - \epsilon)W_{ij}^M) > 0; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

where ϵ is a very small quantity. Thus, the output of an M-unit i is 1 if it receives more input than all other units in its neighborhood, and 0 otherwise, so that active M-units indicate regions of high activity in the feature layer. Each M-unit k then projects to all feature units in the set $\Psi(k)$, which is the set of all feature units in the fan out radius of k.

D. Learning

The weights between feature unit i and feature unit j are strengthened via Hebbian learning as follows:

$$\Delta W_{ij} = \eta y_i y_j \tag{9}$$

The net weight change is a result of both the reinforcement and weight decay. The weight change is

$$W_{ij}^{FF}(t) = min(max(W_{ij}^{FF}(t-1) + \Delta W_{ij}^{FF} - z_i\overline{\eta}, 0), 1) \tag{10}$$

where $\overline{\eta}$ is the rate of decay, and $z_i = 1$ if M-unit k is active, and $i \in \Psi(k)$.

Therefore, every time there is a set of active units in a region, there is both a reinforcement of weights between active units, and a depression of all other weights in the region. When the potentiation rate η is much more than the depression rate, $\overline{\eta}$, the net effect of learning is to make some weights are very strong and the rest very weak.

Usually, weights gradually revert to their baseline values, but once the strengthened weights go above υ and the depressed weights fall to near 0, the weights become fixed and embed the local attractor into the system. The reversion to baseline values occurs as follows:

$$W_{ij}(t) = \begin{cases} W_{ij}(t-1) - \kappa\eta, & \upsilon > W_{ij} > W_{ij}^{bl} + \epsilon; \\ W_{ij}(t-1) + \kappa\overline{\eta}, & W_{ij}^{bl} - \epsilon > W_{ij} > 0; \\ W_{ij}, & \text{otherwise;} \end{cases} \tag{11}$$

where W_{ij}^{bl} are the baseline weights, or the initial weights, before any learning has occured, ϵ is a very small quantity, η is the potentiation rate, and $\overline{\eta}$ is the depression rate. There is a slight reduction in a weight if it is above its baseline weights by more than a certain quantity but below the threshold υ. If it is above υ, the weight is locked. Similarly, depressed weights are locked if they approach 0.

E. Change in firing threshold

A feature of the system is the gradual increase in the firing threshold of a feature unit. As units become habituated to getting a particular amount of input when active, their firing threshold changes over time such that they need this amount of input before they can fire. As a result, a learned attractor needs to be almost fully activated to continue firing, and a partial activation leads to the system forming an attractor in a different region. This is the basis of overlapping attractors in the system.

The equation for the increase the firing threshold of a unit i at time t is:

$$\theta_i^s(t) = \begin{cases} \theta_i^s(t-1) + \xi(gs_i(t) - \theta_i^s(t-1)), & s_i(t) > \delta; \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

where $\theta_i^s(t)$ is the firing threshold for unit i, g is a parameter of the system, and ξ is the rate at which θ_i^s increases. The logic of this function is as follows: $s\theta_i^s(t)$ increases as a fraction g of $s_i(t)$ at time t. However, the actual increase in the threshold is slow, and proportional to the difference between the current value of $\theta_i^s(t)$ and the maximum value it can achieve at time t. Note that $s_i^\tau(t)$ can change only if unit i is active. Also, the increase in threshold starts only once a unit's activity has crossed the limit indicated by δ. Another variable parameter in the system is g. When g is set low, an attractor can be activated based on partial activity, but if g is set high, the system needs to have most of the units in the attractor activated for the system to settle to this attractor.

VII. Simulations and Results

To test the functionality of the system, simulations were run using a 60×60 feature grid, and artificially constructed data. The grid had 3600 feature units that represented 36 features. The data was constucted so that it fell into 6 categories or predefined sets of features that were salient for some context. In addition, there were 2 categories that were not prominent in any one context, but occured sporadically throughout the dataset. We call these *stray categories* for the purposes of this discussion. Each concept occured within a context, and included all features of a category prominent in that context. In addition, concepts could also fall into a stray category. The frequency with which a concept belonged to a stray category was the same as the frequency with which it belonged to a category that was useful for a context.

The simulations were run to test out the following functionality:

1) **Recognition of salient feature sets:** The system had to recognize sets of features that were salient for one or more contexts, and group concepts that were similar to those in the dataset into one of these categories.
2) **Context-dependent learning:** The system had to naturally infer the categories which were useful in a context, while ignoring stray categories.

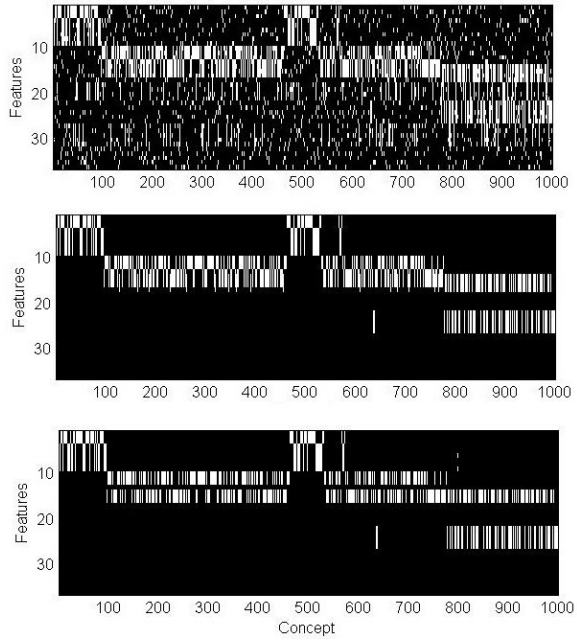

Fig. 3. Clustering of concepts into their salient features - In each of these figures, the x-axis denotes the concepts, or the concept cycles, and the y-axis denotes the features. The top figure shows the concepts that system was tested on after learning. These are different from the training concepts. The actual concepts are noisy and include not just the salient features, but many other features as well. These may not be relevent or prominent in the context. The bottom two figures show the results after learning. Each concept was associated with its inferred category or sets of features. These feature sets are shown here. In the middle graph, $g = 0.9$, and the system tends towards more specific categories. In the bottom graph, $g = 0.5$ and the system tends towards more general categories. In the top figure, concepts from overlapping feature sets are introduced - they fall between features 13-17. Their overlap can be seen clearly in concept cycles 600-1000. A concept switch happens at around concept 800. In the middle figure we see that the concepts were grouped into two separate categories with the differences preserved. On the other hand, in the bottom figure, we see that the concepts fell into one category, defined by the common features alone. Two stray categories that included – feature sets 18-21 and 27-31 respectively – were introduced to the system during learning, and are presented during testing as well. As can be seen in the data, they occur sporadically, but are not prominent in any one context. These two categories were not learned at all.

3) **Context-independent usage:** If a category was seen to be salient in several contexts, it had to be inferred as a single category.
4) **General vs. specific categories:** If there were slight differences, but also significant feature overlaps between different prominently occuring feature sets, the system could either create one category that paid attention to the similarity over different feature sets, or create separate categories that paid attention to the difference among the different features, with the choice being controlled by a system parameter.

The concepts were presented sequentially, by context, as described earlier. The system was trained for 1500 C-cycles (and therefore trained on 1500 different concepts). New concepts that the system had not seen, but which belonged to similar categories were then presented to the system, using the same prototype for the switch in context, and the system output was read out. The results are as shown in Figure 3. As

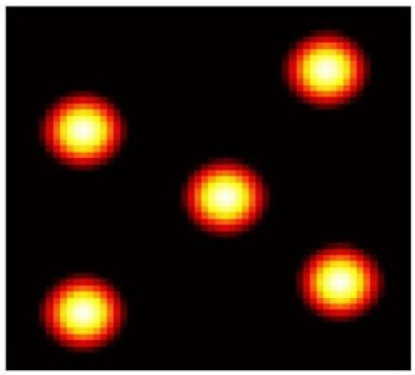

Fig. 4. The initial pattern of weights to five different units on the feature grid. Ligher color indicates a higher excitatory weight; black indicates inhibitory weights.

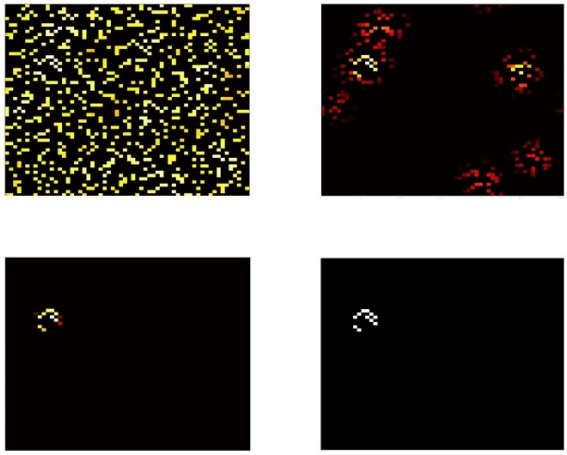

Fig. 5. The feature grid during one concept cycle - The image on the top-left shows the feature grid one recurrent step after the presentation of the concept. The brightness of each pixel on the grid indicates the relative activity level of a feature unit, compared to the activity levels of the other units on the grid. The following images from left to right, and top to bottom show the gradual rise in inhibition, and the convergence of the system to a final attractor. The feature grid on the bottom right shows the units that are learned at convergence.

can be seen in the figure, the system was able to recognize the salient sets of features, infer categories, and group concepts in any of these categories. In addition, the system was able to ignore stray categories. It was also able to choose between learning more general or specific categories.

The next few figures show the actual attractors that were created during the training process. Figure 4 shows the local excitatory weights at baseline for 5 different post-synaptic units in the grid. Figure 5 shows one concept cycle, where the system gradually settles to an attractor. Figure 6 shows the learning of one attractor over many concept cycles, and how attractors are gradually formed in the system. A detailed description of the figures is given in the captions.

VIII. Conclusion

In this paper we have described a self-organized neurodynamical model for category formation, in which the output

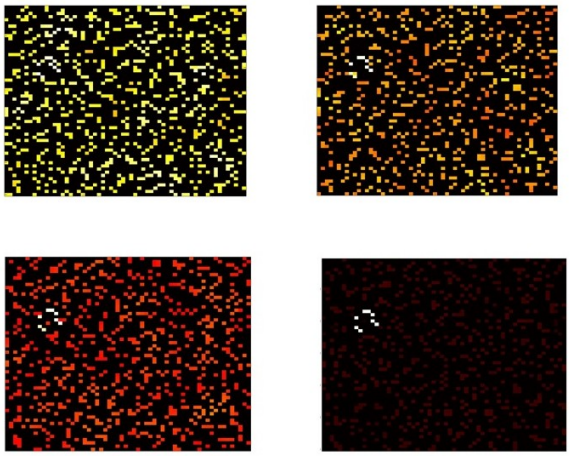

Fig. 6. Learning of one attractor - each figure shows the feature grid one recurrent step after the presentation of the concept. The brightness of each pixel on the grid shows the relative activity level of each feature unit, compared to the activity levels of the other units on the grid. The images from left to right, top to bottom show the system at different stages of learning, in chronological order. In each case, the system settles to the same attractor. As can be seen, the same set of units gradually become stronger over time, and get 'tuned' to a set of features.

emerges from the natural dynamics of the system. We have demonstrated the ability of the system to identify feature sets that are salient in one or more contexts, while ignoring other statistical regularities in the data. Thus the system forms categorical representations that are useful for tasks that the system has actually experienced. Yet, these categories are not specialized for any one task, and may be used in different contexts if necessary. Future directions include the expansion of the model with functional gating, and the testing of the model on real-world data.

ACKNOWLEDGMENT

The authors would like to thank Vince Brown, Simona Doboli, Dan Levine, Paul Paulus and Vaidehi Venkatesan for useful discussions and ideas.

REFERENCES

[1] H.J. Aizenstein, A.W. MacDonald, V.A. Stenger, R.D. Nebes, and J.K. et al Larson. Complementary category learning systems identified using event-related functional mri. *Journal of Cognitive Neuroscience*, 12:977–987, 2000.

[2] F.G. Ashby. A neuropsychological theory of multiple systems in category learning. *Psychological Review*, 105:442, 1998.

[3] F.G. Ashby and W.T. Maddox. Human category learning. *Annual Reviews of Psychology*, 56:149–178, 2005.

[4] H.B. Barlow. Possible principles underlying the transformations of sensory messages. In W. Rosenblith, editor, *Sensory Communications*, pages 217–234. Cambridge: MIT Press, 1963.

[5] L.R. Brooks. Nonanalytic concept formation and memory for instances. In E. Rosch and B.B. Lloyd, editors, *Cognition and Categorization*, pages 169–211. Lawrence Erlbaum, 1978.

[6] J. Duncan. An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews: Neuroscience*, 2:820–829, 2001.

[7] R. Elio and J.R. Anderson. The effects of category generalization and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, 7:397–417, 1981.

[8] J.J. Franks and J.D. Bransford. Abstraction of visual patterns. *Journal of Experimental Psychology*, 90:64–74, 1971.

[9] J.A. Hampton. Testing the prototype theory of concepts. *Journal of Memory and Language*, 34:686–708, 1995.

[10] L.R. Iyer, S. Doboli, A.A. Minai, V.R. Brown, D.S. Levine, and P.B. Paulus. Neural dynamics of idea generation and the effects of priming. *Neural Networks*, 22:674–686, 2009.

[11] L.R. Iyer, A.A. Minai, S. Doboli, V.R. Brown, and P.B. Paulus. Effects of relevant and irrelevant primes on idea generation: A computational model. In *Proceedings of IJCNN 2009*, pages 1380–1387, 2009.

[12] L.R. Iyer, V. Venkatesan, and A.A. Minai. Neurcognitive spotlights:configuring domains for ideation. In *Proceedings of WCCI 2010*, pages 3026–3033, 2010.

[13] J.K. Kruschke. Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44, 1992.

[14] K. Lamberts. Categorization under time pressure. *Journal of Experimental Psychology: General*, 20:161–180, 1995.

[15] K. Lamberts. The time course of categorization. *Journal of Experimental Psychology: Learning, Memory and Categorization*, 24:695–711, 1998.

[16] J.E. Lisman and M.A. Idiart. Storage of 7 +/- 2 short-term memories in oscillatory subcycles. *Science*, 267:1512–1515, 1995.

[17] B.C. Love, D.L. Medin, and T.M. Gureckis. Sustain: a network model of category learning. *Psychological Review*, 111:309–332, 2004.

[18] A.B. Markman and B.H. Ross. Category use and category learning. *Psychological Bulletin*, 129:592–613, 2003.

[19] D.L. Medin and M.M. Schaffer. A context theory of classification learning. *Psychological Review*, 85:207–238, 1978.

[20] E.K. Miller. The prefrontal cortex and cognitive control. *Nature Reviews: Neuroscience*, 1:59–65, 2000.

[21] R.M. Nosofsky. Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13:87–108, 1986.

[22] R.M. Nosofsky. Attention, silimarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57, 1986.

[23] R.M. Nosofsky and T.J. Palmeri. An exemplar based random walk model of speeded classification. *Psychological Review*, 104:266–300, 1997.

[24] R.M. Nosofsky, T.J. Palmeri, and S.C. McKinley. Rule-plus-exception model of classification learning. *Psychological Review*, 101:53–79, 1994.

[25] R.C. O'Reilly and M.J. Frank. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 2:283–328, 2006.

[26] J.M. Pearce. A model for stimulus generation in pavlovian conditioning. *Psychological Review*, 94:61–73, 1987.

[27] J.M. Pearce. Similarity and discrimination: a selective review and a connectionist model. *Psychological Review*, 4:587–607, 1994.

[28] M.I. Posner and S.W. Keele. On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77:353–363, 1968.

[29] P.J. Reber, C.E.L. Stark, and L.R. Squire. Cortical areas supporting category learning identified using functional mri. *Proceedings of the National Academy of Sciences*, 95:747–750, 1998.

[30] S.K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3:382–407, 1972.

[31] E. Rodriguez, N. George, J.-P. Lachaux, J. Martinerie, B. Renault, and F.J. Varela. Perception's shadow: long-distance synchronization of human brain activity. *Nature*, 397:430–433, 1999.

[32] C.A. Seger and E.K. Miller. Category learning in the brain. *Annual Review of Neuroscience*, 33:203–219, 2010.

[33] C.A. Seger, R.A. Poldrack, V. Prabhakaran, M. Zhao, G.H. Glover, and J.D.E. Gabrieli. Hemispheric asymmetries and individual differences in visual concept learning as measured by functional mri. *Neuropsychologia*, 38:1316–1324, 2000.

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

[35] J.D. Smith and J.P. Minda. Distinguishing prototype-based and exemplar-based processes in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28:800–811, 2002.

[36] T. Verguts, E. Ameel, and G. Storms. Measures of similarity in models of categorization. *Memory & Cognition*, 32:379–389, 2004.

[37] E.J. Wisniewski. Concepts and categorization. In H. Pashler and D. Medin, editors, *Steven's Handbook of Experimental Psychology*, pages 467–531. Wiley, 2002.