Self-supervised Learning: A Scheme for Discovery of "Natural" Categories by Single Units

Paul Munro
Institute for Cognitive Science C-015
University of California San Diego
La Jolla, CA 92093

ABSTRACT

Several dynamical systems have been previously proposed to give a neural-like (i.e. connectionist) description of category formation. These typically either involve supervised training (as in Sutton & Barto, 1981; Reilly et al., 1982) or identify dense regions ("clusters") in the stimulus distribution as natural categories (Amari & Takeuchi, 1978; Rumelhart & Zipser, 1985). By combining two existing connectionist-type learning procedures, one supervised and one unsupervised, a hybrid "self-supervised learning" (SSL) mechanism for concept and category learning has been developed. Each unit in the network comes to represent some concept of the order of complexity of a single word; the activity of the unit signals the contribution of its associated concept to the current mental state. A crucial assumption of this approach is that every concept unit (C-unit) receives inputs from two or more information streams. The self-supervised learning process is governed by a data-driven dynamical rule which results in a two-stage learning process. In the first stage, a C-unit becomes selectively responsive to a particular pattern s^{trig} from one of the information streams, ignoring all other patterns in that stream. This is followed by an associative stage in which the unit develops graded response properties to stimulus patterns incident from the other information stream(s). The trigger feature thus becomes a kind of prototype for the concept to be formed by the C-unit. Populations of C-units display interesting representational properties; these are seen to have attributes of both local and distributed representations.

THEORETICAL ANALYSIS

Model architecture

The elements described in this model are labelled C-units (**concept** or **category** units). Each C-unit receives input from two or more (n) groups of afferents (Figure 1), or **input banks**; in principle, n need not be the same for every unit.

In general, indices will follow the convention that superscripts denote the bank and subscripts the component within the bank: afferent j of bank i delivers activity s_j^i via a synapse of strength w_j^i such that a partial response r^i is computed over each bank by the unit in a two-step process consisting of a linear summation followed by a nonlinear "squashing" or "compressing" function:

$$x^i(\mathbf{w}^i, \mathbf{s}^i) = \sum_j w_j^i s_j^i \tag{1}$$

$$r^i = \sigma(x^i) \tag{2}$$

where σ is subject to the condition

$$\sigma(0) = 0 \tag{3}$$

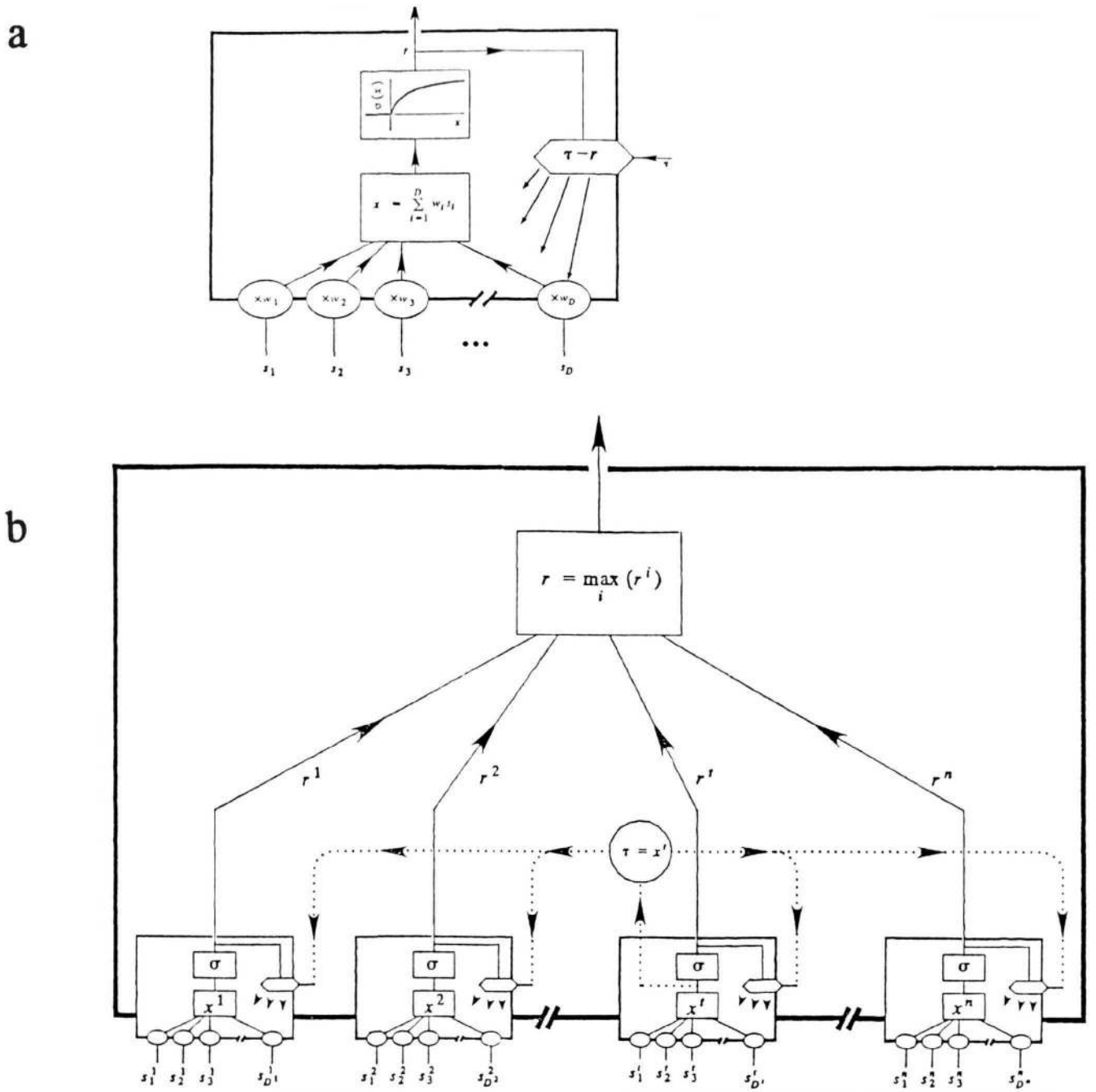

Figure 1. *Information flow in a C-unit.* **a.** An information flow diagram for a single bank, which autonomously follows an algorithm for supervised learning. The stimulus components s_i are weighted by corresponding unit parameters w_i to give a linear activation value x, which is passed to the squashing function σ yielding the unit response r. The value of r is compared with the training signal τ to generate the error value $(\tau - r)$ which is used to adjust the weights according to the rule $\Delta w_i = \alpha(\tau - r)s_i$, where the learning rate α is a small number. **b.** The complete C-unit consists of several input banks (four of these are shown), which each act as a supervised semi-linear unit. Each input bank receives a common training signal τ, but applies the signal to stimulus patterns from different environments. The linear summation stage of one of the banks generates this training signal $(\tau = x^i)$ such that this bank effectively follows the rule for selectivity maximization described in the introduction. The output r of the C-unit is given by a function of the bank responses r^i; in this paper, $r = \max_i(r^i)$.

$$\sigma(x) > x\sigma'(x) \quad \text{for all } x > 0$$

The response r of the unit is a function of the partial sums $r^1 \cdots r^N$. The precise form of this function need not be specified at this point, but it should be nondecreasing in all the r^i; i.e. $\frac{\partial r}{\partial r^i} \geq 0$ for all i. Two cases have been considered – the sum (more generally, an arbitrary linear combination) and the maximum.

The training bank is denoted by the superscript t and is assumed to become selectively responsive to some pattern from its environment E^t. This pattern is the **trigger feature** of the unit and is denoted by s^{trig}. The partial sum and partial response induced by the trigger feature are correspondingly labelled x^{trig} and r^{trig}.

Modification dynamics: the learning rule

The self-supervised learning (SSL) rule is expressed in terms of the time derivatives of the connectivity values w_j^i in terms of the corresponding afferent activity s_j^i, two partial responses (that of the bank to which w_j^i belongs and another that is produced by a special "training bank"), and a variable q that is driven by the training bank's partial response.

$$\Delta w_j^i = \alpha(x^t - q\sigma(x^i))\, s_j^i \tag{4}$$

$$\Delta q = \alpha x^t (x^t - q)$$

where the learning rate α is a small number and the superscript t specifies the *training bank*, such that the partial response x^t "trains" the other partial responses $\{r^i ; i \neq t\}$ to approximate it to the degree that the pattern s^i predicts the pattern to the training bank s^t. The function σ is monotone increasing and satisfies the conditions given by (3).

Final states of the training bank

The SSL equation (4) reduces to the selectivity maximization rule of Bienenstock et al. (1982) along the training bank; the function σ as constrained by Eq. (3) is included to ensure this. For the training bank, equation (4) becomes

$$\begin{aligned} \Delta w_j^t &= \alpha(x^t - q\sigma(x^t))\, s_j^t \\ \Delta q &= \alpha x^t (x^t - q) \end{aligned} \tag{5}$$

The response r^t achieves very high selectivity over the environment E^t. Under the assumption of linear independence within the subenvironments, the training bank attains maximum selectivity; i.e. it responds to exactly one pattern in E^t. Let the chosen pattern, i.e. the trigger stimulus of the training bank, be denoted by s^{trig} and let the corresponding partial sum and partial response [i.e. $\mathbf{w}^t \cdot s^{trig}$] be respectively denoted by x^{trig} and r^{trig}.

Final states of the trained banks

Consider a trained $(i \neq t)$ bank for which the corresponding subenvironment E^i consists of linearly independent patterns. Stable equilibria can then be found by setting the expression (4a) for

Δw_j^i to zero for each pattern s in E^i. If p is the conditional probability that s^{trig} is present on the training bank given s^i at bank i, then for all patterns $s \in E^i$:

$$p(x^{trig} - q\sigma(x^i)) + (1-p)(-q\sigma(x^i)) = 0 \tag{6}$$

If the training bank has reached equilibrium then $x^{trig} = q$ and hence,

$$p = \text{Prob}(s^i = s^{trig} \mid s^i) = \sigma(x^i) \tag{7}$$

REPRESENTATIONS OF STIMULI ACROSS POPULATIONS OF C-UNITS

Up until this point, the description and analysis of SSL has been confined to the single-unit level. While this is appropriate for presentation of the learning mechanism, it is inadequate for understanding certain more global properties, such as those pertaining to the representation of the current state of the world.

If the number of information streams is restricted to just two and all patterns within their subenvironments are equiprobable, then this observation follows concerning the total activity level of the population: *The sum of the unit activities evoked by a given presentation across two information streams decreases with increasing joint probability of the stimulus combination.* That is, the net activity of the population is correlated with the novelty of the stimulus. The relationship between the net activity

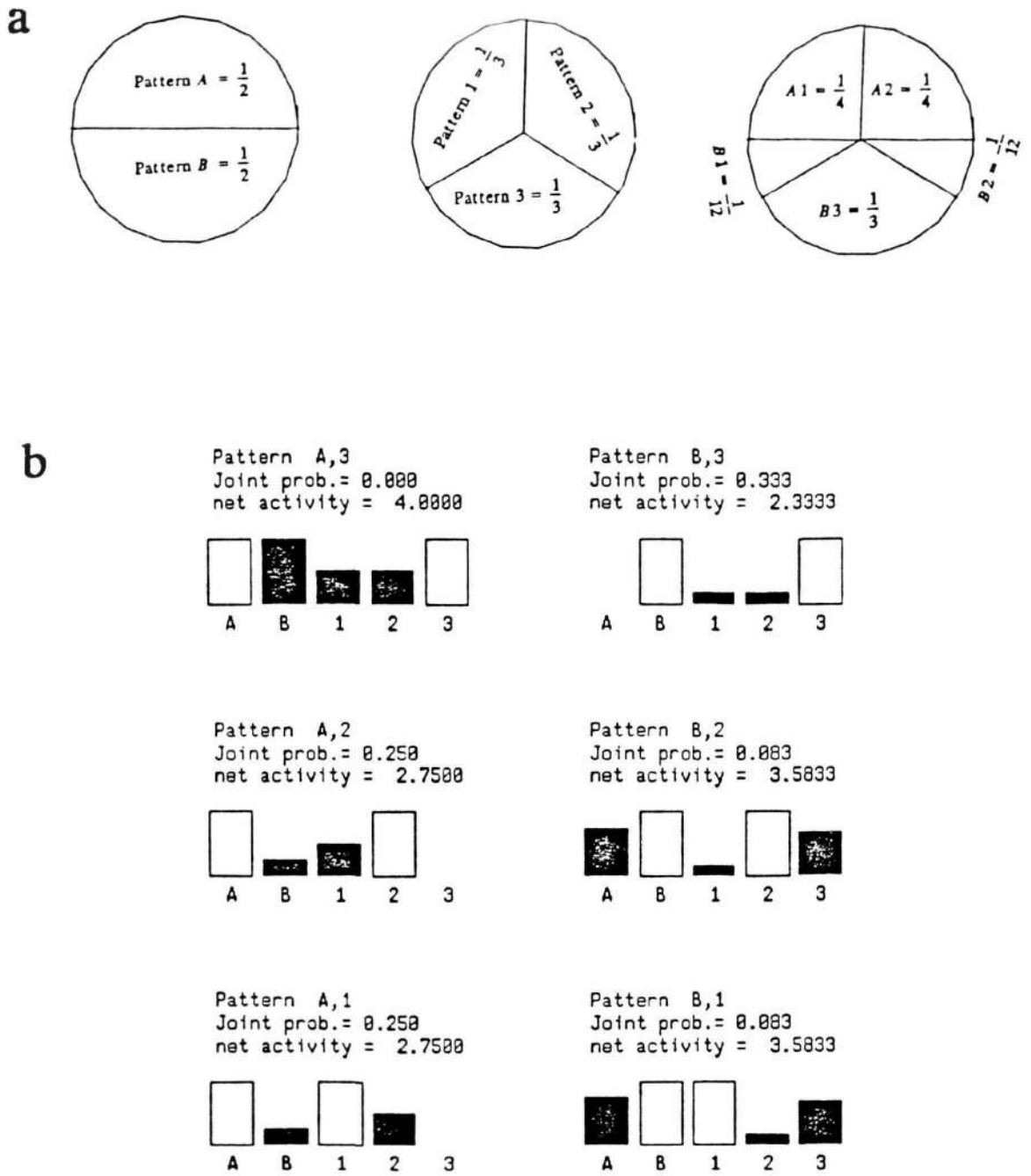

Figure 2. ***A complete representation of a 2x3 environment.*** **a.** The subenvironment E^I consists of the two equiprobable patterns A and B, E^{II} consists of the three equiprobable patterns 1, 2, and 3. Their joint probabilities are shown in the pie chart and the table. **b.** The five subpatterns (A,B,1,2,3) are each the trigger feature for one of the C-units in this minimally complete population. The C-unit representation is shown as an activity pattern over these units for each of the six possible joint patterns, together with the pattern's joint probability and the net activity elicited in the population. For every joint pattern, at least two units correspond to the constituent subpatterns and are thus maximally active. These two components of the activity pattern are shown as unshaded bars, while the "associative" components are shown as shaded bars.

and the joint probability can be shown to be: (by Baye's theorem):

$$A_{TOT}(ij) = 4 - (N_I + N_{II})\, p_{ij} \tag{8}$$

An example system In this example, a population has come to equilibrium with two pattern streams. The respective subenvironments E^I and E^{II} consist respectively of two and three equiprobable patterns: $E^I = \{A,B\}$; $E^{II} = \{1,2,3\}$. Only five $(N_I + N_{II})$ equilibrium states are possible, so for convenience consider a population of just five units, each having converged to a different equilibrium state. Thus the units can be labelled according to the stimulus selected along their training bank. The statistics of the environment and the representations of joint stimuli across the population are described in Figure 2. This simple example illustrates several basic aspects of distributed representations by high-order units. Note that for each pattern, the net activity plus 5 times the joint probability is 4 and hence Eq. (8) is verified.

REFERENCES

Amari, S. & Takeuchi, A. (1978) Mathematical theory on formation of category detecting nerve cells. *Biol. Cybern.* 29:127-136

Bienenstock, E., Cooper, L., and Munro, P. (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* 2:32-48

Reilly, D., Cooper, L., and Elbaum, C. (1982) A neural model for category learning. *Biol. Cybern.* 45:35-41

Rumelhart, D. & Zipser, D. (1985) Feature discovery by competitive learning. *Cognitive Science* 9:75-112

Sutton, R. & Barto, A. (1981) Toward a modern theory of adaptive networks: expectation and prediction. *Psych. Rev.* 88:135-170