

A Biologically Motivated Classifier that Preserves Implicit Relationship Information in Layered Networks

Charles C. Peck

James Kozloski

Guillermo A. Cecchi

A. Ravishankar Rao

IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598

Abstract

A fundamental problem with layered neural networks is the loss of information about the relationships among features in the input space and relationships inferred by higher order classifiers. Information about these relationships is required to solve problems such as discrimination of simultaneously presented objects and discrimination of feature components. We propose a biologically motivated model for a classifier that preserves this information. When composed into classification networks, we show that the classifier propagates and aggregates information about feature relationships. We discuss how the model should be capable of segregating this information for the purpose of object discrimination and aggregating multiple feature components for the purpose of feature component discrimination.

1 Introduction

Classical Artificial Neural Networks (ANNs) are typically arrayed in layers, with each layer sending its responses to higher layers. This arrangement permits each successive layer to respond to increasingly complex combinations of features or attributes in the input space.

These layered networks, however, have fundamental limitations. Consider an array of ANN's that have been trained with visual inputs to realize an orientation-selective, topographically organized map, as in the visual cortex [4]. Each ANN effectively encodes two pieces of information: edge location and edge orientation. These two pieces of information are implicitly related to each other by their shared classifier response. Now consider four additional ANN classifiers that respond to: 1) a vertical edge at any location, 2) horizontal edge at any location, 3) an edge of any orientation in the top half of the image, and 4) an edge of any orientation in the bottom half of the image. Finally, assume that a final ANN classifier exists that responds when a horizontal edge is present in the top half of the image based on the responses of the four classifiers.

If edges can only be horizontal or vertical, this example network will respond properly in each of the four

cases where a single edge is present in the image. When two edges are present, however, an erroneous response can occur. In particular, when a vertical edge is present in the top of the image and a horizontal edge is present in the bottom, both attributes will be simultaneously present and the response will be ambiguous. This is Rosenblatt's "superposition catastrophe" [5].

The root of the superposition catastrophe is that the implicit relationship between an edge's orientation and location is lost when these two attributes are independently classified by the set of four classifiers. Since this lost information is not propagated forward through the network, an ambiguity regarding the sources of attributes can exist at the output.

A related problem is the "component discrimination" problem. This problem involves generating an output response specific to an initial classifier when that classifier contributes to a second classifier giving rise to the final output response. For example, let us modify the previous network such that the horizontal and vertical edge classifiers instead recognize squares and triangles in the image. Let us also modify the output objective such that it is to respond only when a particular edge is active and the edge contributed to the recognition of a square. In this case, implicit information is also lost: the relationship between the edge and the other edges giving rise to the square. This information is subsumed into the "square" classifier. To respond properly, the relationship between the output response and the square classifier must be propagated back to each contributing edge and then forward to the output response. In this manner, the relationship between the edge and the square classifier can be tested by the output response.

2 Solution Requirements

Both the superposition catastrophe and the component discrimination problems result from the loss of information about the relationships among features in the input space or the relationships inferred by higher order classifiers. To solve these problems, therefore, this information must be preserved and made available for classifica-

tion tasks throughout the network.

One approach for preserving this relationship information is to represent it explicitly as a relationship value and couple it with the classifier responses. To solve the superposition catastrophe and component discrimination problems with these values, the following requirements must be met: 1) *Uniqueness*: relationship values must be sufficiently distinct to avoid erroneous relationship interpretations, 2) *Propagation*: relationship values must propagate forward and backward, 3) *Aggregation*: classifiers must take the disparate, but sufficiently similar relationship values of its feedforward and feedback inputs and produce a unified relationship value; and 4) *Selectivity*: relationship values must be used to modulate classifier responses to inputs.

3 A Biologically Motivated Solution

Since the 1989 discovery that neural synchronization correlates with global visual input properties [3], neural oscillation and synchronization have been viewed as a possible means for conveying this implicit relationship information. While both Choe [2] and Seth [6] have modelled neural synchronization, neither effort satisfies the objectives of this paper or operates at the desired level of abstraction. Choe’s work uses a spiking model and it operates at a lower level of abstraction. Seth’s work does not use self-organizing classifiers and may not provide a sufficient basis for classifier learning.

Here, we propose a classifier model that operates on the neocortical minicolumn level of abstraction and uses minicolumn synchronization as the means to manage relationship information. At this level of abstraction, hundreds of neurons are modelled as a single computational system or unit. Furthermore, communication between minicolumns is modelled as an aggregation of the hundreds or thousands of axons that connect them. Thalamo-cortical inputs are modelled as a firing rate, ρ . Cortico-cortical inputs are modelled as pulses up to 12.5ms wide and are represented as a tuple (γ, T) , where γ and T are a pulse’s amplitude and reference time, respectively. While not shown here, this model is based on an analysis of the cortical microcircuit. The model explains observations of neural oscillations as the “chopping” of minicolumn inputs. The unique time required for a minicolumn stimulus to propagate through a specific minicolumn, j , and “chop” subsequent inputs determines the natural period, τ_j of that minicolumn.

In this model, the classification response, ψ_j , of minicolumn j is a weighted sum modulated by the coincidence of each input pulse with the pulse cycle of the minicolumn. This coincidence-based modulation ad-

dresses the selectivity requirement.

$$\psi_j = \sum_{i \in \text{FF}_j} w_{ij} \gamma_i e^{-\left(\frac{T_i - T_j}{2c_1}\right)^2} + \sum_{k \in \text{TH}_j} w_{kj} \rho_k, \quad (1)$$

where T_j is the reference time of the current pulse of minicolumn j , FF_j is the set of feedforward cortical inputs to minicolumn j , γ_i and T_i are the parameters of the current pulse for minicolumn i , w_{ij} is the weight from minicolumn i to j , TH_j is the set of thalamic inputs to minicolumn j , ρ_k is the current firing rate of thalamic relay cell k , w_{kj} is the weight from thalamic relay cell k to minicolumn j , and c_1 is the standard deviation of the Gaussian. The final amplitude response of minicolumn j is: $\gamma_j = \sigma\left(\psi_j, \frac{2.2}{\sqrt{N_{\text{FF}_j} + N_{\text{TH}_j}}}, 0\right)$, where $\sigma(x, \alpha, \beta) = \frac{1}{1 + e^{-\alpha(x - \beta)}}$, the sigmoid function.

Computationally, the aggregation requirement is addressed with synchronization, and synchronization is achieved by adjusting each minicolumn’s reference time, T_j , relative to the reference time’s of other minicolumns using a convex mapping (a modulated weighted sum of reference time differences). The next reference time, T'_j , for a minicolumn’s pulse is computed as follows:

$$T'_j = T_j + \tau_j + \frac{1}{c_2} \sum_{i \in \text{FF}_j \cup \text{FB}_j} (T_i - T_j) w_{ij} \gamma_i e^{-\left(\frac{T_i - T_j}{2c_1}\right)^2}, \quad (2)$$

where FB_j is the set of feedback cortical inputs to minicolumn j and c_2 is a constant corresponding to the maximum contribution from inputs affecting the timing. The use of feedforward and feedback inputs for timing adjustment addresses the propagation requirement. Finally, the uniqueness requirement is addressed by using the natural period, τ_j , as a bias that competes with the pull toward synchronized reference times.

This model provides a basis for learning, but it is independent of the particular method for updating the thalamo-cortical and cortico-cortical weights.

4 Results

Our experimental plan was designed to determine how well the propagation and aggregation requirements are met while simultaneously supporting classification and self-organized learning. We will not explore the uniqueness and selectivity requirements in this paper because those requirements require multiple inputs and an entirely different experimental set up.

The experimental network we used is shown in Figure 1. This network consists of three hypercolumn-like arrays of minicolumns arranged in a hierarchy. At the lowest hierarchical level, two 10x10 networks, denoted

M1L and M1R, each receive feedforward inputs from 2 real values (x, y) in the range $[0,1]$. At the highest hierarchical level, one 10×10 network, denoted M2, receives inputs from the M1L and M1R networks. Each M2 minicolumn receives feedforward inputs from every M1L and M1R minicolumn and each M1L and M1R minicolumn also receives feedback connections from every M2 minicolumn. This connectivity mimics slowly varying projections between hypercolumns. Within each array, every minicolumn makes a center-excitatory/surround-inhibitory pattern of connections with its neighbors.

Our first experiment was to determine whether we could achieve self-organization, both with and without pulsatile inputs. The M1L and M1R results, shown in Figure 1, demonstrate that we were successful in achieving self-organization without pulsatile inputs. Each minicolumn is shaded to indicate the angular location of inputs to which it responds best. The shading key is shown in the bottom of the figure. Similarly, the M2 results demonstrate self-organization based on high dimensionality pulsatile inputs.

Our second set of experiments explored the propagation and synchronization-based aggregation capabilities of our model. The experiments were executed for 30,288 iterations and each stimulus was presented for 24 iterations. Figures 2A–D illustrates activation and synchronization that occurs when the network is presented with a stimulus after self-organization has begun, but not yet completed (completed self-organization is shown in Figure 1). At this stage, we observed that half of the network became sensitive to half of the input space in all three networks. Categories were further subdivided at later stages. Note that synchronization and self-organization are learned concomitantly. Figure 2A shows the classification response of each minicolumn in the three networks. Minicolumns in the bottom half of M1L and M1R are strongly activated by the input, as they are in the right half of M2. The remaining minicolumns are weakly activated.

Figure 2B shows the inter-pulse period for each minicolumn in the three networks. The strongly activated minicolumns, rendered from light gray to white in Figure 2A, share very similar periods. Even minicolumns receiving large numbers of weak inputs, such as the left half of M2, converge to a shared period. Due to the averaging effects in Equation 2, all shared periods derived from large numbers of inputs have similar values. Weakly activated minicolumns in M1L and M1R have diverse periods derived primarily from their unique natural periods. They do not achieve a shared period due to weak feedback from the left half of M2. This diversity also shows the basis for satisfying the uniqueness requirement in our model.

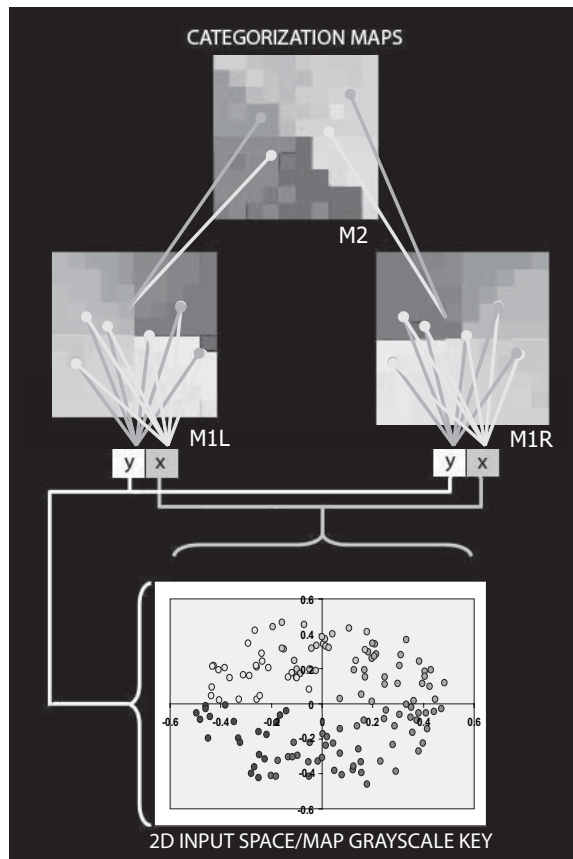


Fig. 1. Experimental Setup. Input space (bottom) and two levels of minicolumn classifiers' (middle, top) preferred-stimulus maps demonstrating self-organization.

The period alone is insufficient to demonstrate synchronization. Figure 2C depicts the pairwise coincidence of pulse reference times. The shades of the lines between two minicolumn conveys the percentage of times their pulse reference times varied by less than 4% during a stimulus. The results demonstrate that minicolumns within strongly activated areas were highly synchronized with each other. In addition, minicolumns in weakly activated areas of M2 were synchronized with each other, but out of phase with the strongly activated M2 minicolumns.

While Figure 2C shows intra-network synchrony, Figure 2D similarly shows inter-network synchrony. In this visualization, the networks are arranged in two planes and all pairs of minicolumns in the network analyzed. The shaded connections show that minicolumns in the strongly activated areas of M1L, M1R, and M2 are synchronized within and *between* their networks. Fi-

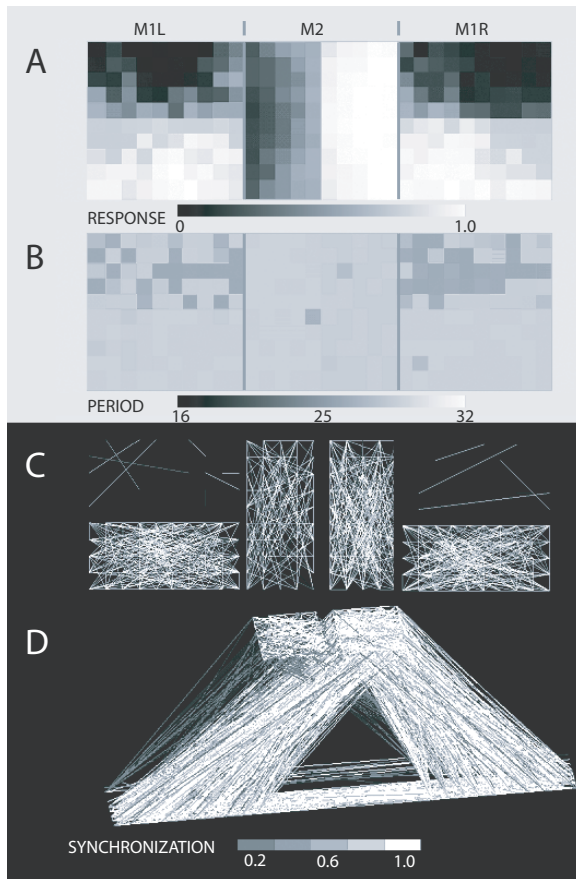


Fig. 2. Results. A. Minicolumn classification response. B. Minicolumn inter-pulse interval in response to a stimulus; C. Minicolumn synchronization within self-organized networks; D. Minicolumn synchronization within and between self-organized networks, dark gray is weak, white is strong.

nally, minicolumns in the weakly activated area of M2 are strongly synchronized with each other and weakly synchronized with some minicolumns in the weakly activated areas of M1L and M1R. This demonstrates that weakly responding minicolumns do not aggregate, which is important for achieving uniqueness.

5 Conclusions

Our results demonstrate that we have solved two key requirements for relationship information preservation: propagation and aggregation of relationship values across a classification network. However, the requirements of selectivity and uniqueness were not fully demonstrated because multiple inputs were not simultaneously presented.

While the forced interdependence between classification and synchronization is not exploited for selectivity and uniqueness in these experiments, the achievement of learning with dynamic inputs is noteworthy and non-trivial. Purely static learning rules are not easily adapted to dynamic inputs [7]. We have solved this problem by learning over those inputs which arrive within a time interval corresponding to a particular phase of an internal oscillation, and by explicitly controlling the relationship between timing and amplitude. We have achieved self-organization, with classification contingent upon synchronization, using both a correlation-based learning approach (Figure 1) and learning rules derived from Kohonen map algorithms [1] (results not shown).

The model also allows for the classification of a feature during a single presentation to become progressively more dependent on only those inputs that are related to the emerging classification. Such a mechanism could allow for more robust classification in the presence of noise, as the mechanism dynamically ignores inputs that do not contribute to the classification. The mechanism should also allow for segregation of nearby features that are derived from different objects, because it allows for classification of a feature to become progressively less a function of those inputs that become more strongly synchronized with other emerging classifications. This selectivity is a requirement for solving Rosenblatt's "superposition catastrophe." By propagating relationship information inferred by classification, this model establishes a basis for grouping features with objects and object discrimination. Experiments using simultaneously presented objects to test the full range of capabilities of our model are ongoing.

References

- [1] Bednar, J. A. & Miikkulainen, R. (2003) *Neurocomputing*, 52-54:473-480.
- [2] Choe, Y. & Mikkulainen, R. (2004) *Biol Cybern.*, 90(2):75-88.
- [3] Gray C. M. & Singer W. (1989) *Proc Natl Acad Sci U S A* 86: 1698-1702.
- [4] Grinvald, A. & Lieke, E. & Frostig, R. D. & Gilbert, C. D. & Weisel, T. N. (1986) *Nature*, 324(6095):361-4
- [5] Rosenblatt, F., (1962) *Principles of Neurodynamics: Perception and the Theory of Brain Mechanisms*. Washington: Spartan Books
- [6] Seth, A.K. & McKinsty, J.L. & Edelman, G.M. & Krichmar, J.L. (2004) *Cerebral Cortex* May 13 Advanced access.
- [7] Song S., & Abbott L. F. (2001) *Neuron* Oct 25;32(2):339-50.