66

SELF-TEACHING THROUGH CORRELATED INPUT

Virginia R. de Sa
Dana H. Ballard

Dept. of Computer Science, University of Rochester, Rochester, NY 14627

ABSTRACT

Previous work has shown that competitive learning coupled with a top-down teaching signal can produce compact invariant representations. In this paper we show that such a teaching signal can be derived internally from correlations between input patterns to two or more converging processing streams with feedback. Such correlations arise naturally from the structure present in natural environments. We demonstrate this process on two small but computationally difficult problems. We hypothesize that the correlations between and within sensory systems enable the learning of invariant properties.

66.1 INTRODUCTION

Unsupervised neural network learning algorithms, which are limited to classifying patterns based only on the similarity of their input representations, are unable to form position invariant and other image invariant representations. Previous work [3] has shown, using the architecture in Figure 66.2a), that competitive learning coupled with a top-down teaching signal can learn these task-relevant representations of the input patterns as shown for the XOR problem in Figure 66.1.

For this problem the network must learn to separate the two sets of input patterns, {(-1,+1),(+1,-1)} and {(-1,-1),(+1,+1)}. Figure 66.1a) shows graphically the weights of neurons in the first layer of an unsupervised competitive network. One weight is shared between two of the closest patterns, but as these patterns are from different classes, future layers cannot separate the classes. Figure 66.1b) shows the effect of adding another weight from a teaching signal. The feedback signal adds an extra dimension to the weight space and influences the relative distances between patterns in the new augmented space allowing an appropriate representation (see Figure 66.1c).

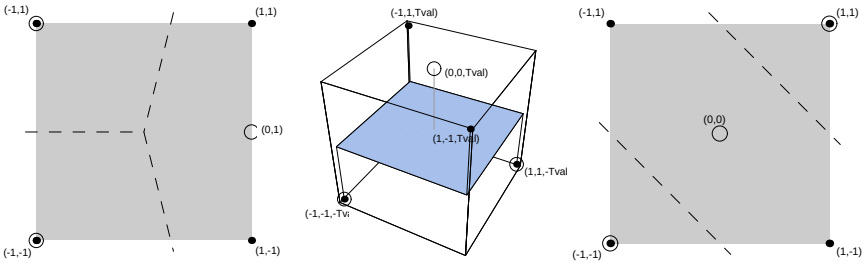

Figure 66.1 Input patterns are represented by the small dark dots, weights are represented by the larger open circles, and the dashed lines represent the partition of the input space among the neurons.
a) The patterns and weights for the XOR problem with 3 hidden-layer neurons and no teach input. The solution has allocated one weight to be shared between two of the closest patterns. The other two weights are allocated to the other patterns.
b) The patterns and weights for the XOR problem with 3 hidden-layer neurons and a teach input. The third dimension comes from the feedback from the output layer. The solution has allocated one weight to be shared between two neurons with the same feedback. The other two weights are allocated to the other patterns.
c) The result of removing the Teach input after training as in b).

The algorithm's dependence on an externally derived teaching signal, however, is unsatisfactory from a biological perspective. In this paper we demonstrate that an internal teaching signal can be derived from correlations between input patterns to several networks. That is, a collection of semantically correlated input patterns can collectively teach themselves [1].

Consider an infant learning to recognize his parents. The infant receives many visually dissimilar views of his mother and father as well as many voice samples of different words in different tones from both parents, yet he must learn to recognize his parents' faces and voices. We assume that he cannot simply memorize every visual and auditory instance (corresponding for example to one hidden layer neuron per pattern) and that he is not born with appropriate "mother's face" type feature detectors but instead must develop appropriate invariant representations. In addition he has no external teaching signal classifying each instance of visual and auditory data.

What he can use is a benevolent environment in which the sensations of the two (and more) modalities are correlated. An image of his mother's face usually appears with an instance of his mother's voice and likewise for his father's face and voice [2]. We hypothesize that these correlations between and within sensory systems can drive the development of appropriate representations in each modality[3].

[1] This is similar to [5] and [2] except that we do not have one modality train the others—all modalities cooperate to train themselves.

[2] There are also temporal correlations within modalities, but our algorithm currently does not make use of this information.

[3] This idea is put forth in [6], however as in [3] the simultaneous development of the sensory representation and association is not demonstrated.

66.2 ARCHITECTURE AND ALGORITHM

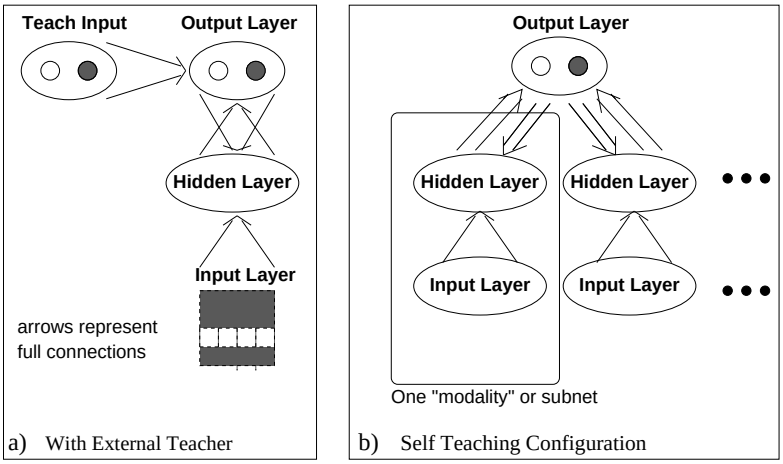

Figure 66.2 a) The architecture used in [3]. The hidden and output layers are competitive networks. The arrows represent full connections in the directions shown
b) The modified self-teaching architecture. The teaching input has been replaced by another sub-network. Together the networks train each other according to the correlations in their patterns

We demonstrate these ideas using the architecture shown in Figure 66.2b). Incoming sensations to each sensory modality are represented as activation in an input layer. They project (through separate hidden layers) to a common output layer. This common output layer projects back to all hidden layers and serves as the shared extra dimensions just as the externally taught output layer does in Figure 66.2a).

The network operates exactly as that in [3] [4] except that the output activation is determined by the activations of the hidden layers. Weight updates are similar except that the weights of the hidden layer neurons active on the upward pass (with no active feedback) are moved away from the input patterns[5].

These updates allow neurons responding to temporally correlated patterns to develop similar connections to and from the output neurons while at the same time encouraging neurons to respond to patterns receiving the same feedback. The key to the algorithm is that all patterns within one class in in each modality occur with a unique distribution of patterns in the other modalities, and these distributions serve as a type of stochastic teaching signal for each other. This problem is non-trivial though, as we are not simply

[4] Each layer within each modality net performs a competitive winner-take-all calculation, meaning that only one neuron is active in each layer at each time. In the output layer the winning neuron is given activation `Tval` (all other winning neurons are given activation 1).

[5] This is similar to LVQ[4] and can be seen as an abstracted form of the BCM [1] weight update rule in that during the upward pass (where activations are not as strong due to lack of input on the descending connections) weight updates are anti-Hebbian and during the downward pass they are Hebbian.

learning a mapping from static hidden layer representations to the output neurons, but this mapping is changing the hidden layer representations themselves. The correct representations must develop in concert with the correct connections from the hidden to the output layers.

66.3 RESULTS

This approach solves two small but computationally difficult problems. The XOR problem was tested for networks of two, three, four and five "sensory modalities". At random one of the four input patterns from the two sets of patterns {(-1,+1),(+1,-1)} and {(-1,-1),(+1,+1)} was presented to the first network. For the original tests, the other networks received, with equal probability, either the same pattern or the other pattern from the same set (Note that in this case all modalities receive inputs from the same pattern space).

The networks were able to learn the correct representation strictly through correlations between the two inputs. Figure 66.3a) shows the performance for different combinations of output activation strength (`Tval`) and number of modalities. For these graphs correct performance was defined as the development of a correct mapping for each input modality, thus higher modality nets start at a disadvantage in that there are more input nets to develop correctly. The figure shows that as the value of `Tval` becomes more useful [6] the nets with more modalities perform better. That is, more networks provide a more reliable output and when this output is allowed to be effective more modalities can offset their more stringent requirements [7].

Subsequent trials tested the effect of relaxing the correlations on the two-modality network. For these trials the second network received a pattern from the correct class with various probabilities [8]. Figure 66.3b) shows the graceful degradation in performance for correct correlation probabilities between 100 and 50%.

We also tested the algorithm on the task of distinguishing horizontal and vertical lines. Again both "modalities" were given inputs from the same pattern class (There were 8 possible lines in a 4×4 pixel array.) and the goal was to learn to activate a different output neuron for horizontal and vertical lines. A network of two modalities was consistently able to learn the lines problem with 6 hidden layer neurons in both subnets. A more efficient, with only 4 neurons per sub-net hidden layer, encoding was achieved by adding a third sub-net.

66.4 SUMMARY

In summary, experiments show that association between different input streams can

[6] higher values result in hidden neurons ignoring the input and lower values are not strong enough to counteract the input similarities

[7] These figures were made for simulations of 400 time steps. For longer simulations all the plots migrate up — given sufficient time all the tested combinations gave 100% correct performance over 400 trials.

[8] This corresponds to occasionally hearing father's voice while seeing mother's face.

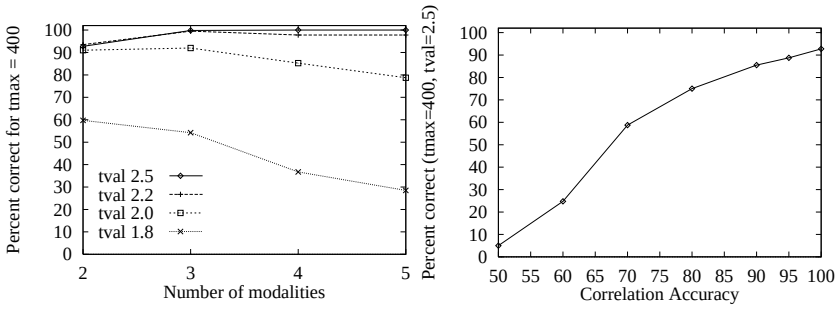

Figure 66.3 a) The effect of output activation strength Tval on convergence accuracy (percent correct performance as calculated from 400 random trials) for different numbers of modalities in the XOR problem.
b) The effect of correlation accuracy on convergence accuracy for a 2-modality net. X% of the time the second modality received a pattern from the same class and 100-X% a pattern from the other class.

influence the representations of all the streams. This problem is hard in that the network must detect correlations and simultaneously change its encoding which looks for the correlations. Thus, the whole system must bootstrap itself to achieve both the right representations and the right association.

Acknowledgements

We would like to thank Randal Nelson and Robert Jacobs for helpful conversations. This work was supported by a grant from the Human Frontier Science Program and a Canadian NSERC 1967 Science and Engineering Scholarship.

REFERENCES

[1] E. L. Bienenstock, L. N. Cooper, and P. W. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.*, 2:32–48, 1982.

[2] G. A. Carpenter, S. Grossberg, and J. H. Reynolds. Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, 4:565–588, 1991.

[3] V. R. de Sa and D. H. Ballard. Top-down teaching enables task-relevant classification with competitive learning. In *IJCNN International Joint Conference on Neural Networks*, volume 3, pages III–364—III–371, 1992.

[4] T. Kohonen, G. Barna, and R. Chrisley. Statistical pattern recognition with neural networks: Benchmarking studies. In *Proc. IEEE Int. Conf. on Neural Networks, ICNN-88*, volume 1, pages I–61—I–68, 1988.

[5] P. Munro. Self-supervised learning of concepts by single units and "weakly local" representations. Technical report, School of Library and Information Science, University of Pittsburgh.

[6] E. Rolls. The representation and storage of information in neuronal networks in the primate cerebral cortex and hippocampus. In Durbin, Miall, and Mitchison, editors, *The Computing Neuron*, chapter 8, pages 125–159. Addison-Wesley, 1989.