

Engram Decay in Artificial Neural Networks

Howard Copland and Tim Hendtlass
Centre for Intelligent Systems
School of Biophysical Sciences and Electrical Engineering
Swinburne University of Technology
PO Box 218 Hawthorn 3122 Australia
E-mail: hxc@brain.physics.swin.oz.au and tim@bsee.swin.edu.au

1. Abstract

In this paper the process of forgetting - termed "engram decay" - in artificial neural networks is examined for three classes of networks: the self organizing map, fuzzy logic adaptive resonance theory and maximally connected back propagation networks. All networks were trained to categorize three varieties of iris. How iris categories are forgotten is shown to be strongly related to the distribution of iris categories in feature space.

2. Introduction

The precise mechanisms underlying memory in biological neural networks have yet to be determined, though studies indicate that it is stored in a distributed - frequently termed *holographic* - state. Karl Pribram[1] elaborated on this theme and developed the hypothesis of the "holographic brain". Most hypotheses concerning memory involve *engrams* or "memory traces", which are considered to be the biophysical manifestation of memory in the brain.

Given this paradigm for biological memory, of no less importance is the process by which information is forgotten. Forgetting is an integral function of any biological information system. An emerging neurophysiological hypothesis is that the brain functions as the sum of many interacting resonant systems, and in this light forgetting may be considered damping of selected elements of one or more of those systems. This process of forgetting may be examined to some extent via artificial neural networks (ANNs), though work in this area has been very limited to date.

As models of the biological neural networks, ANNs cannot be said to be emulating neural function as neural function is inherently chemical, but instead working to the same processing paradigm. Consequently, the simulations are intended as plausible representations, as opposed to actual mimicry, of the processes involved.

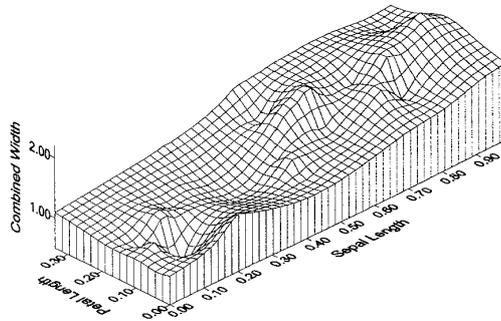
The practical application of knowing the dynamics of memory decay may lie with problems that require long training times - many hours to days. Should overtraining of a network occur, it would be useful to know how long to train such a network on a biased data set in order to "regeneralize" it to optimum rather than start over altogether.

3. Quantifying the Data Set

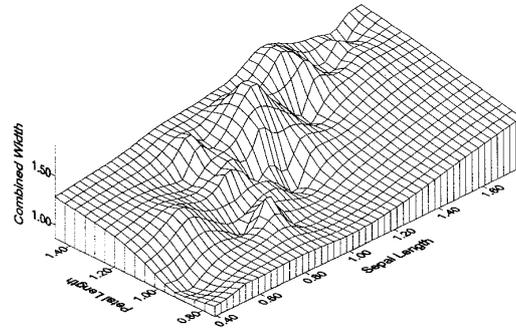
The data set consisted of the sepal and petal lengths and widths for 140 individual irises. Irises are sorted to one of three categories - Setosa, Versicolour and Virginica - based upon the length and width of sepals and petals. An analysis of the data set is necessary as *what* a network forgets has bearing on *how* a network forgets.

The technique of multidimensional scaling was employed to determine how similar the three iris categories were to each other. Vector summation of the two "width" axes - these possessing the least variance - allowed for three dimensional representation of the four dimensional data, and here the similarities between the categories are evident.

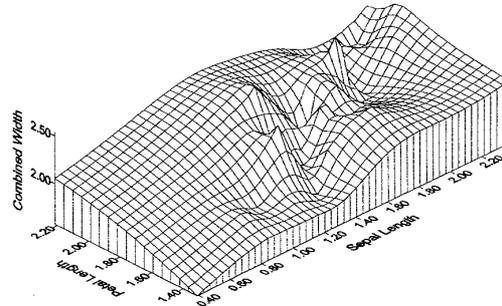
Iris Category A



Iris Category B



Iris Category C



This corresponds to a scaled feature space separation of the category centroids as follows, the units are arbitrary:

<i>Category Centroids</i>	<i>Distances</i>
Distance of Category A from Category B	1.65
Distance of Category A from Category C	2.83
Distance of Category B from Category C	1.04

The four dimensional (S4 or hyper) volume that each category occupies in feature space gives an indication as to the distribution of the data. A hypercubic arrangement for category feature spaces is assumed. Again, the units are arbitrary.

<i>Categories</i>	<i>Hypervolumes</i>
Category A	0.30 units ⁴
Category B	0.72 units ⁴
Category C	1.60 units ⁴

The extent to which the categories intersect is a further measure for determining the nature of the data. Category A intersects Category B and Category C in one plane only: the sepal length by sepal width plane. Category B and Category C intersect on 3 planes generated by axes of sepal length, sepal width and petal length.

<i>Planar/Volume Intersection</i>	<i>Congruency(%)</i>
Category A intersects Category B (2D)	31.65%
Category B intersects Category A (2D)	35.00%
Category A intersects Category C (2D)	37.31%
Category C intersects Category A (2D)	30.06%
Category B intersects Category C (3D)	14.36%
Category C intersects Category B (3D)	8.66%

While the planes of intersection are the same for any two categories, the *relative* extent of the intersection is different as the categories occupy different sized hypervolumes. This may seem misleading in

that volume congruencies - congruencies being the extent to which the regions are accordant - are lower than the planar congruencies, but a volume congruency involving Category A is zero. The planar congruencies for Category B and Category C in the sepal length by sepal width plane are considerably higher than those involving Category A:

Category B intersects Category C (2D)	64.38%
Category C intersects Category D (2D)	58.27%

Indeed, the sepal lengths for Category B are entirely contained within the range for the sepal lengths of category C. That is, sepal length is not a distinguishing feature between Category B and Category C.

4. Results

Examined were backpropagation networks[2], self-organizing maps[3] and adaptive resonance theory networks[4]. Once the network had learned to categorize the irises, a single category of iris was removed from the training set, and the training of the network was continued. The performance of the network was then recorded at intervals.

4.1 Backpropagation

The backprop networks were of 4-5-3, 4-4-3 and 4-3-3 architectures, and trained for 50,000 cycles. The hidden layer and output layer learning coefficients were 0.3 and 0.15 respectively. The network performance prior to the removal of the category was the benchmark against which later network performance was measured.

All networks categorized the irises with 100% accuracy; network size seemed of only marginal importance, with negligible differences in the confidence indices ($\Delta C.I. = 0.04$) of the smallest and largest networks. The minimum value for any correct categorization was ~ 0.9 for all networks, for an ideal of one (1). The confidence index indicates the extent to which a network has uniquely identified an input. It is determined by taking the highest output value and subtracting from it the second highest value. Incorrectly classified examples produce a negative confidence index.

For all networks, following the removal of a category, there was a rapid decline in performance after the first 20,000 training cycles. This decline was on the order of 50% for Category B and Category C, and on the order of 5% for Category A. A significantly lower rate of decay for the next 100,000 cycles followed, where upon a new 'equilibrium' position is reached within the network and the performance decays no further. This new equilibrium position is function of network size - the larger the network, the lesser the decay.

As this is happening, the minimum output values of the two categories still being presented oscillate, both increasing and decreasing as they compete for newly freed resources. At a given point, this competition will end and the new equilibrium will be established. This process can be considered in terms of a contracting feature space around the centroid of the disabled category. The minimum drops much more rapidly than the mean average, as it is the outlying examples of a category which are potentially closer to another category, so on the edge of a category's feature space.

The decay of Category A in all networks was considerably more gradual than the decays of Category B or Category C. The differences between Category A and the other categories are greater and so more 'memorable' to the network. Category B and Category C are separated by more subtle differences and it is these which decay rapidly.

The decay can be described by the simple exponential functions:

$$a \cdot \exp[-x' \cdot (b+cx)] \quad \text{cycles} < 20,000 \quad (1)$$

$$1 - \exp[-x' \cdot (a+bx')/(x'+c)] \quad \text{cycles} > 20,000 \quad (2)$$

where x is cycles and x' is cycles/1000;
and, coefficients a , b , c are data set dependent.

Equation 1 is a good description of the decay curve, but only in the range of the rapid decay itself. Extending the function beyond the first 20,000 cycles in order to find the "memory half-life" is invalid as a new

equilibrium point is reached; the networks tend toward stability. For network behaviour beyond 20,000 cycles, the second function provides a far better description.

4.2 Self-Organizing Maps

SOMs of architectures 10x10, 12x8 and 12x7, all with a learning coefficient of 0.06, were trained for 4,200 cycles. Low confidence indices - 0.038, 0.086 and 0.088 respectively - for the trained network indicate the SOM's categories were quite different to those of human convention. Network size seems not be especially important, though the largest network did perform least well.

As with backprop, the SOM did categorize Category A well away from Category B or Category C. Category C was almost as equally well defined as Category A, with Category B apparently caught in the middle of these, and consequently ill-defined by the networks with mean performance averages under 0.5. This is likely due to the large congruency between Category B and Category C.

When an entire category is removed, the features which are unique to that category are not updated on the map. Where no updating occurs, the network becomes static. Consequently, the unique features are not forgotten as in backprop, but they are not reinforced. As two other categories are still being presented to the map, and these possess features common to themselves and also the removed category, the map does change. Those congruencies existing between the removed category and either of the remaining categories will, additionally, no longer be reinforced. This process of forgetting occurs rapidly, within an additional 1,000 training cycles in all cases.

When the SOMs performance on the removed category is tested, the only identifiers for the category are the static regions, and what remains of the congruencies of all *three* categories. These static regions dictate the minimum level of recall.

4.3 Fuzzy Adaptive Resonance Theory Networks

FzART networks were trained for 560 cycles, all categorizing with 100% accuracy. All had four nodes in the F2 layer, varying only in the memory stability factor, termed α . This factor is a component of the long time memory, relating to the extent the "synapses" are influenced by a given input; valid values are between zero and one, with zero making the long term memory unalterable. The choice of four nodes - representing four internal categories or templates - was established as the minimum required to correctly classify the iris data set. Four internal templates were required for the three iris categories as Category C required two templates due to the large volume of features space occupied by Category C irises.

Forgetting in a FzART network relies on instability and the peculiarities of the adaptive resonance architecture. A stable network does not forget whereas an unstable one will. The minimum level of instability, as measured by α , was found to be 0.2 in order for any memory decay to occur. The α values evaluated were 0.1, 0.2, 0.3 and 0.4.

Uniquely, it was not the removed category which was necessarily forgotten. Indeed, no network misclassified a Category A or Category C iris even after 10,000 presentations of a data set with the respective category absent.

Where Category A was removed, Category B irises were misclassified as Category C. Where Category B was removed, Category B irises were again, as expected, misclassified as Category C. Where Category C was removed, no misclassifications occurred.

This counter-intuitive result derives in part from the learning mechanism employed by network and in part from the implementation of fuzzy logic. When an input is presented, it is classified and the feature space assigned to that category is enlarged all the way to the current input example, the template for the category being updated is made to look more like the input example.

The impact of fuzziness on the performance of the network is best demonstrated by example. Removing Category A causes Category B to be misclassified as the network is employing fuzzy logic. A trained network may assume, for a given example, the following values:

	<i>Category A</i>	<i>Category B</i>	<i>Category C</i>
<i>Example n</i>	0.17	0.44	0.39

This input example would be classified (correctly) as a Category B iris. The "winner-take-all" approach of the output drives this 0.44 to 1 and the other outputs to zero. When Category A is disabled, however, the classification may look more like this:

	<i>Category A</i>	<i>Category B</i>	<i>Category C</i>
<i>Example n</i>	-	0.45	0.55

The iris is incorrectly classified as a Category C iris, and so influences Category C template(s). It is this modified template that causes more Category B examples to be misclassified as Category C, causing the feature space for Category C to expand.

Misclassification of this type is of greater magnitude than the misclassification that occurs when Category B is removed all together. This is due to Category C templates becoming more like the Category B irises still being presented in the first instance and thus expanding the Category C feature space beyond the bounds of Category C examples. When no Category B irises influence the Category C templates, as happens when Category B is removed, the misclassifications are derived only from the Category B irises congruent with Category C.

When Category C is removed, Category A and Category B templates are sufficiently dissimilar to the Category C template(s) so as not effect the networks' recall of Category C.

5. Conclusions

In terms of biological plausibility, of the networks examined, it was the maximally connected backpropagation networks that showed what was considered the most biological behaviour with respect to forgetting: a gradual, near exponential, decay of the engram. Yet of the networks examined, backpropagation is frequently considered the least biologically plausible overall.

Its biological similarity lies in that the engram is distributed across all the nodes in the network. In SOMs it is the similarity of the input and exemplar vectors that determine a node's activity; FzART templates are similarly stored in a single node of the F2 layer. Neither technique involves true distribution of memory.

Exponential decay of memory may prove to be feature of holographic systems: a property of the paradigm.[5]. Such decay has proven to be case with optical holograms, on which the paradigm is based.

In all instances and regardless of network type, forgetting was found to be very dependent on the distribution of the data in feature space.

6. References

1. Pribram, K., Brain and Perception, L. Erlbaum and Associates Inc., U.S., 1991.
2. Hendtlass, T., Private Communication, 1994.
3. Neural Computing: A Technology Handbook for Professional II Plus and Neuralworks Explorer, Neuralware Inc., Pittsburgh, 1993, pp299-303.
4. *ibid*, pp158-166.
5. Psaltis, D., Brady, D., Hsu K., "Learning in Optical Neural Networks", Parallel Processing in Neural Systems and Computers, Elsevier Science Publishers, North Holland, 1990, pp543-551.