Managing Interference Between Prior and Later Learning

L. Andrew Coward[1], Tamás D. Gedeon[1], and Uditha Ratanayake[1]

[1] Department of Computer Science, Australian National University,
Canberra, ACT 0200, Australia
[2] Dept. of Electrical and Computer Engineering
The Open University of Sri Lanka, PO BOX 21, Nawala, Nugegoda, Sri Lanka

1. Introduction

A barrier to complex learning in artificial neural networks is catastrophic interference between later learning and earlier learning [French 1999]. There have been numerous attempts to address this issue [Robins 1995] but success has been limited as the complexity of the problem addressed by the network increases. A critical issue for learning complex combinations of capabilities is maintaining adequate meanings for information generated by one part of a network but used for different purposes in many other parts of the network [Coward 2001]. Avoiding interference in these terms depends on maintaining information meanings, as learning changes information. This paper describes an investigation of the interference between prior and later learning in a connectionist architecture called the recommendation architecture in which explicit management of the maintenance of operational meanings is possible.

2. Implementation of the Recommendation Architecture

Properties of recommendation architecture systems have been investigated in learning problems such as telecommunications network management and document classification [Coward et al 2001; Ratnayake et al 2002]. There is a separation in the recommendation architecture between clustering which defines and detects conditions in the information available to the system and competition which associates different combinations of conditions with different behaviours. The device used in clustering is illustrated in figure 1A. The weights of inputs which define conditions to this device are binary. The device detects activity in a set of regular inputs, and produces an output if a subset larger than a threshold is present. The set is defined by converting active provisional inputs to regular if the number of active regular inputs is less than the threshold, the total of active regular and provisional inputs is above the threshold, a signal exciting the recording of conditions is present, and no signal inhibiting the recording of conditions is present. The effect of this algorithm is that once a device has produced an output in response to the activity of a specific set of inputs, it will always in the future produce an output in response to the activity of the same set. This permanence is in contrast with perceptron type algorithms standard in ANNs in which the adjustments to individual weights mean that a device may not respond to an exact repetition of a condition which earlier generated a response.

Condition recording devices are arranged in layers. Inputs to devices in the first layer are system inputs, condition defining inputs to devices in the second layer are from first layer devices and so on. The complexity of conditions detected increases through the layers, where complexity is defined as the total number of system inputs which contribute to a condition. There are additional inputs from devices in specific modules to condition recording devices in specific other modules. These additional inputs are from special

purpose devices which detect the level of activity of condition recording devices in the source module, and excite or inhibit changes to conditions detected by their target device.

A modular hierarchy is superimposed on the device layers, in which modules have specific functional roles. The first level of module is made up of devices in an area on one layer. The second level is a column made up of corresponding areas on a sequence of four layers. The third level is an array of parallel columns. The functions of these modules can be understood by consideration of two columns in the same array as illustrated in figure 1B. In each column, layer 4 is a single special purpose device that produces a binary output which is 1 if any of the devices in layer 3 of the column are active, otherwise 0. The layer four device is connected to devices in competition.

Initially, all the condition recording devices in a column have randomly selected provisional condition defining inputs. Condition defining connectivity between layers is only within columns. Columns are initiated one at a time, no other column is initiated until the previous column is generating a layer 4 output in response to a significant percentage of input states. Biases are placed on the random selection process for provisional inputs in favour of groups of inputs which have tended to be active in the same system input state in the past. Once some columns have been initiated, groups are defined with a bias in favour of inputs which have tended to be active at the same time in input states for which no previously initiated column has activity above a threshold level in its layer 1.

A high level of input activity into layer 1 of a column excites initiation of a column and inhibits initiation of any other column. The first input state will therefore initiate a column. This initiation means that condition recording devices in layers 1 to 3 will define regular conditions until there is activity in layer 3, at which point further definition will be inhibited. The next input state could have one of several results. One is that regular conditions are detected in all levels 1 through 3 and the column produces an output. Recording of conditions will be inhibited. A second is that there is activity in layer 2 but not in layer 3. In this situation, additional conditions will be recorded in layers 1 through 3 until layer 3 activity inhibits further recording. The third is that there is activity in layer 1. Such activity will inhibit the initiation of another column. When multiple columns have been initiated, there is another possibility. There could be activity in layer 2 of several columns, but inhibition between columns limits recording to the column with the highest proportion of active layer 2 devices.

The effect of a series of input states is that columns are defined each detecting a somewhat different portfolio of conditions. The array definition process thus results in compression of a large number of input characteristics which discriminate weakly between different categories into a much smaller number of portfolios which discriminate much more strongly between the categories. Increase in discrimination means that an individual characteristic can occur in some instances of N different categories, but a column is active in response to instances of n different categories, where in general $n << N$.

Category identification achieved by competition associating different combinations of portfolio outputs with different categories is therefore easier than with system inputs. Each portfolio output is assigned a weight in favour of each category, and weights are adjusted until high integrity identifications are achieved. This weight adjustment can be guided in two ways. In supervised feedback the competition component corresponding with the correct category is identified. In consequence feedback it is indicated whether the category identification by the system is correct or incorrect.

If the learning process in competition does not converge, the lack of convergence is an indication that the columns provide inadequate discrimination between categories. Such a lack of convergence would be demonstrated, for example, if at different times when the same set of columns was active and generated a category identification, the identification was sometimes correct and sometimes incorrect. This situation triggers generation of additional columns which provide additional discrimination.

The values of the device algorithm and modular hierarchy are that they result in column outputs in response to every input state; they result in consistent column outputs in the

sense that once a column has generated an output in response to an input state it will always produce an output in response to an identical input state; they manage condition recording to prevent excessive recording; and they compress a large input space into a smaller output space with better category discrimination.

In the terminology of unsupervised learning in ANNs, columns correspond with heuristically defined clusters [Gokcay and Principe, 2002]. The reference vector for a column is implicit in the conditions recorded in layer 2, and is evolved whenever an input state is added to the column. The critical difference from unsupervised learning is that if an input state contains a condition which has been assigned to a cluster, an exact repetition of that condition in another input state will always be detected by the same cluster.

Competition devices in figure 1B have excitatory inputs from the special purpose devices in layer 4, a maximum of one from each such device. These inputs have continuously variable weights, and devices produce outputs proportional to the sum of the input weights of all active inputs. The category corresponding with the device with the largest output is the selected identification. When the category corresponding with a competition device is indicated by supervision, an input to that device is established from any layer 4 devices in columns with output activity which does not already have such a connection. This new input is given a standard weight. If the total of the weights of all active inputs to the device corresponding with the target behaviour is less than the total for any other device, the weights of the currently active inputs into the target device are all increased by the same proportion until the output from the target device is the largest.

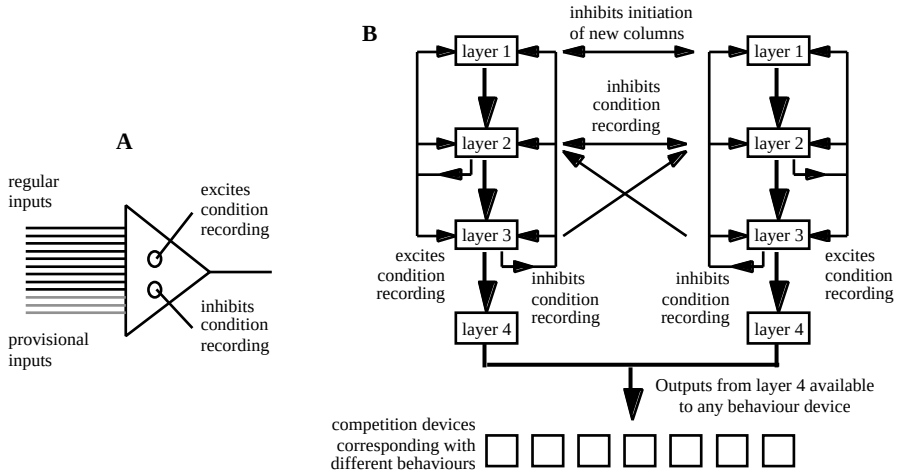

Fig. 1. A. Condition recording device. B**.** Modules in clustering, and outputs to competition

In consequence learning, if the category is correct, the weights of active inputs to the competition device with the largest output are increased by 5% of the standard weight. If the category is incorrect, the weights of those active inputs are decreased by 10% of the standard weight. Under no feedback conditions, the identity of the category is compared with the identity generated by the system, but no feedback is provided.

Interference between later and earlier learning could occur in a number of ways. One is that changes to column portfolio definitions during later learning might dilute the discrimination of the portfolio. This dilution is limited because conditions must be similar to the existing portfolio to be added and if different conditions are needed they will be will be assigned to a new column. Dilution is also limited because as the portfolio expands, the

number of devices which must be active in layer 2 for recording to occur increases. Hence the degree of difference from the existing portfolio under which addition of conditions is allowed decreases with learning. Finally, because recognition of a category is recommended by the presence of a number of portfolios, the effect of change to one portfolio will be limited.

Any column will typically target a number of different category components in competition. The second way in which interference between later and earlier learning could occur is when the weights assigned as a result of supervision to inputs to a new category device are partially for columns which also have weights into a previously learned category. For some combinations of column outputs in response to an instance of the old category the total weight into the new category could exceed that into the old category. This dilution will be limited by the degree of discrimination between the categories provided by the columns, and also can be controlled by consequence feedback in response to incorrect identifications.

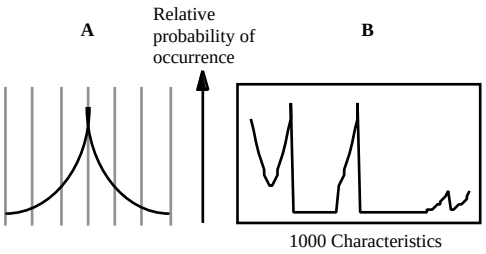

Figure 2 Statistical generation of category instances. A. Probability distribution shape divided into six segments. One of the six is selected at random to define the relative probability for each 100 characteristic block in a category, plus a 40% chance of selecting zero probability. B. Relative probability of occurrence of characteristics in an example category.

3. Testing the Interference Between Early and Later Learning

The test scenario to investigate the effects of later learning on prior learning was that the system was presented with a series of category instances. The presence or absence of 1000 characteristics could be discriminated in each instance, and the information available to the system from each object was therefore equivalent to a 1000 bit binary vector.

Instances were defined by the set of characteristics which they possessed. Sixty different categories were defined by probability distributions for the occurrence of characteristics in instances of the category. The probability distribution for a category was created by random selection of a distribution shape for each block of 100 characteristics. The six possible shapes were the six segments of the curve illustrated in figure 2A. In each selection there was an equal chance of each shape being selected plus a 40% chance of a zero probability of characteristics in the block appearing in instances of the category. An example distribution is illustrated in figure 2B. All categories contained characteristics which only appeared with very low probability and are similar to noise.

1000 objects were constructed for each category. The construction process was random selection of characteristics from a set with the appropriate distribution for the category. The number of characteristics in a object varied in the range 70 - 180. No two objects in the same category were identical. Individual characteristics occurred in instances of between 15 and 33 categories out of the 60, with an average of 26.1 categories or 43.4% of the total. This percentage is a measure of the ability of individual characteristics to provide discrimination between categories of inputs.

The adjustable parameters in clustering in these experiments were the same as those described in Coward [2001]. In the process for the heuristic definition of columns, there were periods in which sequences of input vectors were presented to the system (wake periods) separated by sleep periods in which provisional connectivity was configured in new and existing columns in preparation for the next wake period. Early experiences were limited to instances of 50 categories, later experiences included instances of 60 categories.

In supervised feedback, the identity of the competition component corresponding with the presented category was provided. In consequence feedback, a correct or incorrect indication was provided in response to the category identified. Under test conditions, no feedback was provided. After column definition in response to the first 50 categories, supervised feedback was provided for between 1 and 32 instances of those categories. At this point (P1) identification accuracy was determined for 5 instances of each of the 50 categories. Columns were then extended by experience of the full 60 categories followed by supervised feedback for between 1 and 32 instances of the new 10 categories. At this point (P2) identification accuracy was determined for 5 instances of each of the 60 categories. An alternative learning profile was that consequence feedback on the original 50 categories was provided when supervised feedback was provided on the new categories. At this point (P3) identification accuracy was determined for 5 instances of each of the 60 categories. In the entire process including column definition, supervised and consequence feedback, and testing phases, all instances presented to the network were different.

Four sets of columns were generated in parallel by the same presentations. These sets each divided up the input similarity space in somewhat different ways. The purpose of this generation of multiple sets was to investigate whether addition of sets of columns could be a strategy for improving the overall operational discrimination of clustering outputs.

4. Test Results

An array of columns contained on average about 23 columns, including about one column added during learning of the extra 10 categories. One column does not in general correlate unambiguously with one category. On average a column generated an output in response to some instances of about a quarter of the categories. Since on average a characteristic appeared in some instances of about half of the categories, this represents an improvement in operational discrimination by a factor of 2. Category identification accuracy for one array after supervised learning of the original 50 categories was 45%. Accuracy improved to 75% with two sets, 85% with three sets, and over 90% with four sets. Learning capability can thus be improved by additional sets of columns in the same input space. The rest of the results reported here are for four sets of columns. The number of columns in four sets averaged 93, compression by over 10 from the 1000 characteristic input space. Clustering thus achieved both compression and higher operational discrimination.

The first learning process achieved a 91.6% accuracy for instances from 50 categories. In general the variation in recognition accuracy using different object instances but with the same experience generating clustering was ±0.5%. The variation in recognition accuracy for different clustering experiences was ±2.5%.

Table 1 shows the accuracy with which instances of the first 50 categories were identified following the different learning processes. The recognition accuracy for the extra ten categories was greater than 90% at points P2 and P3. For many of the errors in identification, the second choice identification was correct. In other words, the competition device corresponding with the correct identification had the second highest output activity. For example, in table 1 for 16 supervised instances, 88.4% of the instances of the first 50 categories were correctly identified before learning the additional 10 categories (point P1), but only 44.4% of instances after the additional learning (point P2). However, for 32.4% of the instances at P2, the first identification was incorrect but the

second choice was correct. This type of information robustness accounts for the way in which identification accuracy was restored to 81.6% if simple consequence feedback on identifications of the original 50 categories was provided during supervised learning of the new 10 categories (point P3).

Table 1 demonstrates that subsequent learning of an extra 10 categories has an effect on prior learning of the 50 categories, but the effect is a gradual degradation not catastrophic destruction. The gradual degradation is the result of slight broadening of the similarity definition for some columns, and of clusters with weights in competition indicating some of the original categories gaining large weights in favour of new categories. Both of these effects can be limited by providing simple consequence feedback for identifications of the original categories during the period in which supervised feedback is provided for the new categories.

Number of instances of each category used during supervised training	Final recognition correctness measured on 5 instances of each of 50 original categories at end of process:		
	P1	P2	P3
1	67.6%	26.4%	38.0%
2	72.4%	31.2%	51.6%
4	75.6%	32.8%	70.8%
8	84.4%	40.4%	77.6%
16	88.4%	44.4%	81.6%
32	91.6%	44.8%	79.6%

Table 1 Recognition accuracy of original 50 categories before and after learning of the additional 10 categories, with and without consequence feedback on identifications of the original categories during supervised feedback on the additional categories. Results are averages for five experiments.

4 Conclusions

It has been demonstrated that using a condition recording device algorithm in an architecture called the recommendation architecture, interference between later and prior learning is not catastrophic, and can be managed by providing limited consequence feedback when supervised learning is occurring and by increasing the number of condition portfolios in a controlled fashion.

5 References

Coward, L. A., Gedeon, T. D. and Kenworthy, W. (2001). Application of the Recommendation Architecture to Telecommunications Network Management. *Int. Journal of Neural Systems* 11(4) 323-327.

Coward, L. A. (2001). The Recommendation Architecture: lessons from the design of large scale electronic systems for cognitive science. *Journal of Cognitive Systems Research* 2(2), 111-156

French, R. M. (1999). Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Science* 3(4), 128-135.

Gokcay, E. and Principe, J. C. (2002). Information Theoretic Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(2) 158-169.

Ratnayake, U. and Gedeon, T. D. (2002). Application of the Recommendation Architecture for Discovering Associative Similarities in Documents. Proceedings of the 9th International Conference on Neural Information Processing.

Robins, A. (1995). Catastrophic Forgetting, rehearsal, and pseudorehearsal. Connection Science 7, 123 - 146.