Monograph:
Downs, J. and Harrison, R.F. (1997) A Survey of Neural Network Models Based Upon Adaptive Resonance Theory. Research Report. ACSE Research Report 664 . Department of Automatic Control and Systems Engineering

A Survey of Neural Network Models Based Upon Adaptive Resonance Theory

Joseph Downs, Robert F Harrison

Department of Automatic Control and Systems Engineering

The University of Sheffield

Research Report 664

27 January 1997

Abstract

This report provides an introductory overview of Adaptive Resonance Theory (ART) including the motivation behind the theory. It then describes various neural network models based upon ART. Such models are categorised as to whether they utilise unsupervised or supervised learning. Notable examples of the former that are described include ART 1, ART 2 and fuzzy ART. Examples of the latter include ARTMAP, fuzzy ARTMAP and fusion ARTMAP.

Correspondence Address
R.F. Harrison
Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street
Sheffield, S1 3JD
United Kingdom
Telephone: +44 (0)114 2225139
Facsimile: +44 (0)114 2780409
E-mail: r.f.harrison@sheffield.ac.uk

1 Introduction

Neural network learning schemes can be broadly distinguished by whether they are supervised or unsupervised. The former type of network requires the presence of a teaching stimulus external to the network which provides the correct output during teaching that will be associated with each input pattern. Probably the best-known example of this type of learning is the back-propagation model (Rumelhart, Hinton and Williams, 1986) where a multi-layer network is trained by adjusting its weights based upon differences between the actual and desired output of the network. These differences are then propagated down each layer of the network from output to input units.

In contrast unsupervised network models do not employ external error-correction. Instead, they learn to detect for themselves regularities in the input patterns presented to them. Such self-organisation results in the development of feature detectors which can reliably recognise important similarities in the input stimuli. The existence of such detectors allows the network to categorise subsequent novel input patterns. Of course, unlike teacher-directed pattern classification, these categories have not been externally provided, but were derived by the network itself in response to regularities amongst its training inputs.

Perhaps the best-known example of this paradigm is competitive learning (Rumelhart and Zipser, 1985). This is a feedforward architecture with purely excitatory connections between layers of nodes. Within an individual layer, units are grouped into clusters so that intra-layer connections occur only between units within the same cluster. All such connections are inhibitory, thus within each cluster a competition occurs between units to become the most highly activated. The intention is that each cluster will code for a different feature in the input stimuli, while each unit within a cluster codes for a different value of that feature.

In the most simple competitive learning mechanisms, only the most highly activated unit in a cluster is allowed to learn from each training example. This scheme is thus called "winner-take-all". Learning is achieved by a unit transferring weight from inactive to active input lines.

2. The Stability-Plasticity Dilemma

Unfortunately there is a significant drawback with competitive learning theory (and indeed many other neural network learning schemes), identified by Grossberg (1987). This is that the network's responses are temporally unstable such that during training previously learned useful codes can be over-written by (possibly irrelevant) new codes. And hence a particular feature detector may have an entirely different response to the same input pattern each time it is presented owing to perturbation in the memory system caused by different intervening input patterns.

Grossberg calls this problem the stability-plasticity dilemma: "How can a learning system be designed to remain plastic, or adaptive, in response to significant events, and yet remain stable in response to irrelevant events?" (Carpenter and Grossberg, 1988, p.77).

The typical solution to this dilemma, widespread throughout the connectionist community, is an unsatisfactory one. This is to terminate all further learning by the network once an acceptable (and externally-determined) level of performance is achieved on the training set. Thus a dichotomy is enforced between learning about the environment and responding to it, where the network is only able to display plasticity during the initial learning phase.

This solution is unacceptable from the viewpoint of both the psychologist and the technologist. A psychologist would find the rigid separation between learning and responding highly

unrealistic for both actual neurons and at the higher level of observed human behaviour. A technologist must be concerned that in any moderately complex domain the training set will be unlikely to encapsulate the full variety of the performance environment. This of course means that the system's behaviour may not always be adequate for all contingencies in the domain.

3 Adaptive Resonance Theory

The stability-plasticity dilemma is however convincingly solved by adaptive resonance theory (ART). ART was developed from competitive learning theory (c.f. Grossberg, 1987; Carpenter and Grossberg, 1988). It supplements the bottom-up (or data-driven) processing of competitive learning with a mechanism for top-down (or expectation-driven) processing and an orienting mechanism to distinguish between familiar and unfamiliar inputs.

Processing in an ART network intuitively occurs in the following general manner. A stimulus pattern I activates a layer of input nodes F1. These in turn activate a node Na (representing a feature or category, or more accurately a category prototype) at the category representation layer F2, above F1. (This is bottom-up processing.) The code at Na is then transferred down to F1 for comparison with the input pattern. (This is top-down processing.)

If an acceptable match between I and Na is found, the bottom-up input pattern and its top-down expectation are said to be in a resonant state of mutual reinforcement. Learning then occurs such that the F2 node Na winning the competition for I is able to adjust its weights (long-term memory or LTM).

If however a poor match occurs, a reset signal is sent to F2 by the orienting mechanism which suppresses Na and allows another node, Nb, to be triggered. Matching then repeats as before with this node, and so on. If no top-down expectation (corresponding to an existing learned code) can be found adequately to match to I then a previously uncommitted node at F2 will be selected. The learning of a new category will thus occur. This formation of new categories can continue until the memory's capacity is reached.

Processing in ART is able to distinguish between familiar and unfamiliar events. The attentional subsystem, consisting of top-down processing together with the bottom-up competitive learning between F1 and F2, deals with processing of familiar events (inputs), where old categories are recognised and refined. The separate orienting subsystem handles unfamiliar events (i.e. it is effectively a novelty detector) and resets the main attentional subsystem when they occur, allowing the formation of new categories without over-writing earlier ones. Thus adaptive resonance theory is able to resolve the stability-plasticity dilemma without having to terminate learning until the memory becomes full.

It can also be seen that ART employs direct access to codes for familiar input patterns, supported by search for matches to less familiar (or entirely novel) inputs. This is an appealingly efficient way of organising memory.

Another feature of ART, interesting from a psychological viewpoint, is the capability for priming effects. It is a well-known finding of cognitive psychology that activation of the memory of a particular category can facilitate recognition of exemplars of that category, while hindering recognition of exemplars of other, unrelated categories. The expectation-driven matching in ART seems to provide exactly this.

Two important further aspects of ART have been omitted to date however:

1) Vigilance. The formation of categories at the F2 layer will be critically determined by the degree of mismatching that is tolerated between top-down expectations and bottom-up input

patterns. This can be set within ART by means of a vigilance parameter, ρ, that is responsible for activating the orienting mechanism when an unacceptable mismatch is deemed to have occurred. A low vigilance setting allows a relatively high degree of mismatch to be tolerated, resulting in the formation of very general (or coarse-grained) feature detection nodes. A high vigilance setting conversely creates narrow (or fine-grained) categories as only small mismatches are now allowable. Therefore correspondingly more reset bursts are sent by the orienting mechanism to the F2 layer in this case.

The vigilance parameter can further be used to allow a degree of teacher-directed learning within ART's basic self-organising capabilities. In particular, the environment can provide negative feedback should ART erroneously categorise an input pattern. This will result in increased vigilance within the network, forcing finer-grained discriminations to be made. (Conversely, the teacher may signal that a particular input pattern should be easy to recognise and vigilance may be reduced.)

2) Attentional Gain Control and the 2/3 Rule. In order for input nodes at F1 to be triggered correctly it is necessary for there to be a means to distinguish between bottom-up and top-down activation. This is achieved by an activation source called the attentional gain control which connects to all nodes in the input layer. It is thus a third source of activation to these nodes, in addition to attentional priming from the category representation nodes and the actual input patterns from the environment.

The intention is that the attentional gain control and attentional priming cannot simultaneously activate input nodes. During bottom-up competitive learning, the attentional gain control provides non-specific activation to all nodes in F1. In matching expectations to inputs, the top-down priming mechanism suppresses the gain control, so that only the priming signal and the input pattern act across the nodes.

These modes lead to the 2/3 rule that an input layer node can only transmit activation if it is itself activated by two of its three possible sources of stimulation. This rule is fundamental to the stabilisation of code learning in ART. It's violation can have effects of psychological interest however. For example, activation by attentional priming alone is akin to the network misinterpreting internal states as perceptions (as in dreams or hallucinations).

4 Unsupervised Learning: ART 1, ART 2, ART 3 and Fuzzy ART

The description of adaptive resonance theory given to date is most appropriate to a particular class of network models called ART 1 (Carpenter and Grossberg, 1987a). ART 1 is only able to process binary input patterns. Moreover, it has been shown to exhibit less stable learning than was originally claimed, in some circumstances (Ryan and Winter, 1987). This problem will be described further in section 6.

ART 2 (Carpenter and Grossberg, 1987b) is able to process analogue or binary input patterns. The extension of the model to direct processing of analogue inputs obviously allows for continuous-valued or grey-scale patterns as stimuli. However, it also requires the inputs to be contrast-enhanced to suppress noise prior to category recognition. This is achieved in ART 2 by the addition of a pre-processing layer F0 and a more complicated F1 layer in comparison to ART 1. This latter layer contains several processing layers and gain control systems so that a bottom-up stimulus no longer directly matches the top-down expectations but must first undergo preliminary processing where positive feedback loops enhance salient features and suppress noise. Typically the F1 layer of ART 2 contains three such levels; the bottom layer

normalises the bottom-up input stimuli, the top layer normalises the top-down expectations, and matching between the normalised patterns occurs at the middle layer.

Since there is no direct matching of inputs to expectations in ART 2 a modified (and weaker) version of the 2/3 rule is used in which a node only remains active during matching if it receives a significant amount of top-down activation. This is still sufficient however for code learning to self-stabilise. It should also be noted that although the ART 2 architecture is generally more complex than ART 1, the learning mechanisms are somewhat simpler since the same learning rule is used for both the top-down and bottom-up weights.

Carpenter, Grossberg and Rosen (1991a) introduce ART 2-A, an emulation of ART 2 that performs at processing speeds two or three orders of magnitude faster than the original. ART 2-A is biased more strongly than ART 2 to immediate selection of an uncommitted node following a reset rather than a search of committed nodes. The model emulates ART 2 under fast learning and intermediate learning conditions. Fast learning allows the LTM weights to be fully recoded by presentation of a single input pattern. This allows category prototypes to be formed with fewer input stimuli. However, such prototypes are highly vulnerable to noise in a particular pattern obliterating important features of the prototype. Intermediate learning consequently allows only partial recoding of the LTM weights by a single input presentation, while retaining the rapid commitment to a particular node. This reduces the susceptibility to noise while maintaining rapid processing speeds.

ART 2 is also modified in Carpenter and Grossberg (1990), this time so as to allow ART 2 modules to be organised into a hierarchical cascade. In this arrangement the output of an F2 layer serves as the input to a higher F1 layer. This results in the formation of increasingly abstract categories. Such an organisation, called ART 3, requires the connections between layers to be modified so as that each layer has the same type of links. Thus there is no longer the strict distinction between separate F1 (input) and F2 (output) layers, since each layer in ART 3 can subsume both functions.

ART 3 also requires a new search mechanism for resets due to mismatching. This is achieved by means of a biologically realistic model of chemical transmission at the synapses. The modified search mechanism allows ART 3 to support distributed coding of categories. The "winner-take-all" rule used in earlier ART models resulted in a localist representation, with each different category prototype being coded by a single F2 node. (Grossberg refers to these as "maximally compressed" codes.) This method simplifies the competitive learning mechanism, and thus eases comprehension of the network dynamics. However it also loses the graceful degradation of performance shown by neural networks with distributed representations. However such representations could not be used in earlier ART models due to the inability to search a F2 layer with distributed coding following a reset signal.

The ART models described so far all have their underlying mechanics based upon operations from traditional set theory. Thus category membership in these models is an all-or-nothing affair. Carpenter, Grossberg and Rosen (1991b) introduce fuzzy ART, a model that replaces the intersection operator of traditional set theory used in ART 1 operations with its equivalent from fuzzy set theory, the minimum operator.

This results in fuzzy ART being able to process binary or analogue input patterns like ART 2. Moreover, the input patterns in fuzzy ART can represent crisp or fuzzy feature sets. In the latter case, the presence of a feature is valued between 0 and 1 to indicate the extent to which it is present. Of course in the case of crisp-featured, binary input patterns, fuzzy ART performs identically to ART 1, and is also capable of processing crisp-featured, analogue

inputs similarly to ART 2. (Indeed fuzzy ART has now largely superseded ART 2 since it provides equivalent processing power using a much simpler and more elegant model.)

Carpenter, Grossberg and Rosen (1991b) also describes a pre-processing scheme for the inputs to fuzzy ART called complement coding. They strongly recommend that this scheme be used in conjunction with fuzzy ART since the model is otherwise prone to problems of category proliferation.

5 Supervised Learning: ARTMAP, fuzzy ARTMAP and variants

Adaptive resonance theory evolved from competitive learning. Accordingly ART 1, ART 2, ART 3 and fuzzy ART are all essentially unsupervised learning systems. Obviously, learning without the presence of a teacher is an essential element of a psychologically plausible theory of intelligence. It is also useful in many application domains where environmental feedback may be unavailable or at least unreliable. However it is sometimes necessary for a specific mapping to be learned by a neural network. Unsupervised learning cannot guarantee that any such categorisation will arise. It is entirely possible that unintended categorisations will be learned as the network discerns a different structure in its inputs to that it was intended to detect. Indeed the formation of feature detectors is dependent on the order in which particular inputs occur. (And in pure competitive learning schemes it is also dependent on the initial random weighting of the network.)

Thus it can be seen that a general learning system, while having the capability of learning autonomously, should also be able to make use of tuition where it is available. More recent models of adaptive resonance theory have therefore addressed the issue of teacher-directed learning.

ARTMAP (Carpenter, Grossberg and Reynolds, 1991) is one such model. Its architecture consists of two standard ART 1 modules, ARTa and ARTb, which respectively receive the input pattern and its associated teaching stimulus. Linking ARTa and ARTb is a third module, called the map field, which associates the recognition categories output from ARTa with those output from ARTb. It can thus be seen that ARTMAP does not directly associate an input pattern to its training pattern. Rather the input and the teaching stimulus each evoke a compressed recognition category at the F2 layers of ARTa and ARTb respectively. It is these recognition codes that are associated at the map field, hence generalised associations are formed.

Successful formation of associations is critically dependent upon internal control of the map field gain control and orienting subsystems. In particular the map field orienting subsystem is used to control the vigilance parameter, ρ_a, of the ARTa module. (This is known as match-tracking.) Thus in the event of a mismatch between two codes at the map field, vigilance is increased so as to trigger a search for a new category code at ARTa that will be associated with the code from ARTb.

ARTMAP is claimed to possess a number of desirable properties; learning is rapid and very accurate, fine-grained discriminations are possible even when a rare input pattern must be distinguished from some other highly-similar and frequently occurring pattern, and the system is capable of real-time on-line learning.

However, since ARTMAP is based upon ART 1 modules it is restricted to the processing of binary input patterns. Fuzzy ARTMAP (Carpenter, Grossberg et al, 1992; Carpenter and Grossberg, 1992) rectifies this limitation. Its general architecture is much the same as ARTMAP with the exception that the ARTa and ARTb modules are no longer ART 1 systems

but fuzzy ART systems. This provides the capability for processing binary or analogue patterns which may have crisp or fuzzy features. The performance of fuzzy ARTMAP has been shown to compare favourably with back-propagation and genetic algorithms on a number of benchmark classification tasks.

A number of extensions have now been made to the basic fuzzy ARTMAP model. One such area of research has been the extraction of symbolic rules from a trained fuzzy ARTMAP network (Carpenter and Tan, 1993; Tan, 1994). Such rule extraction is a simpler process for fuzzy ARTMAP than for backpropagation due to the localist representation of categories and lack of hidden units. Indeed, in essence, each committed ARTa node represents a symbolic rule whose antecedent is the category prototype weights and whose consequent is the associated ARTb category (pointed at via the map field). However, fuzzy ARTMAP often generates large numbers of categories with highly variable weights, resulting in rules that are difficult to comprehend both individually and collectively. To overcome these difficulties, rule extraction involves two "pre-processing" stages, pruning and quantization.

Pruning involves the deletion of category nodes deemed to be of low utility for categorisation (which typically represent rare but unimportant cases). Pruning is guided by the calculation of a confidence factor for each category, based upon a category's volume of usage and accuracy at category prediction. With some large databases, pruning actually improves fuzzy ARTMAP performance on novel test data, by curing the network's over-specification on items in the training set that are not of general importance.

Quantization involves fixing the continuous-valued weights of the categories that remain after pruning to a set number of feature level values. This improves the comprehensibility of the final rules by ensuring uniformity of weight values.

One limitation of the fuzzy ARTMAP model is that it has no provision for dealing elegantly with missing data items. A variant of fuzzy ARTMAP has been developed, therefore, called fusion ARTMAP (Asfour et al., 1993) which is claimed to provide graceful degradation of performance in the face of missing "chunks" of data. Fusion ARTMAP is basically a generalisation of fuzzy ARTMAP, consisting of a number of separate modules each of which is dedicated to independent pattern classification of different types of inputs (or sensor readings). A global classifier then fuses data from the separate modules to reach a final category decision. If information from one or more modules is not available, a category decision can still be made using the information from the other channels.

The ART-EMAP model (Carpenter and Ross, 1993) is another variation of the fuzzy ARTMAP model which is intended to provide distributed representation of category prototypes (akin to the ART 3 model for unsupervised learning). However, like ART 3, it requires a more complicated model to achieve this aim.

6. Advantages of Adaptive Resonance Theory

Models based upon adaptive resonance theory offer a number of advantages over other forms of neural network. From the technologist's viewpoint, they are particularly appealing in that the solution to the stability-plasticity dilemma provided by ART models allows them to be utilised in a wider variety of applications than, for instance, the feedforward networks. For example, ART models do not require a well-bounded and stable input environment, performing robustly under noisy conditions in a nonstationary world. Moreover, since they do not require arbitrary cessation of learning, they can be trained on-line. And, of course, the

unsupervised ART models do not require the presence of teaching stimuli to ensure category formation.

The similarity-based learning in ART also offers advantages over error-based learning schemes such as back-propagation. Training times are much faster, since stable categories can be formed from a single input presentation without requiring averaging over a number of events. Thus learning can occur in real-time. In addition, it allows fine-grained discrimination between categories which might otherwise be erroneously merged by the averaging process. Furthermore, there are no constraints (such a linear independence) on the form of the input patterns, nor is learning subject to the problem of local minima found in backpropagation.

ART is also appealing from the psychologist's viewpoint since it is also intended as theory of cognition in addition to its uses in specific applications. Many aspects of adaptive resonance theory are biologically or psychologically realistic, (with the notable exception of the non-distributed category coding found in all models except ART 3). Indeed, Grossberg regards pattern recognition in ART as an exemplification of general cognitive processes such as hypothesis discovery and testing, attention, search, classification and learning. C.f. Grossberg (1988) and Carpenter and Grossberg (1991) for numerous examples of the use of ART to explain neurophysiological and psychological data.

7. Limitations and Modifications of Adaptive Resonance Theory

7.1 Variants of ART 1

A number of other researchers outside of Grossberg's group have identified limitations and improvements on various adaptive resonance models. Most such work has concentrated upon ART 1, since this is, of course, the most amenable to formal analysis.

For example, Georgiopoulos, Heileman and Huang have studied the learning properties of ART 1 in detail. Georgiopoulos, Heileman and Huang (1990) prove that during fast learning for a special class of ART 1 models the recognition code self-stabilises after at most λ list presentations for patterns of size λ. (A size λ pattern having λ bits set to λ and all others set to 0.) This result holds regardless of the ordering of the patterns in the list for presentation purposes, and thus provides a useful general upper bound for the amount of training needed by this ART 1 sub-class during fast learning.

Georgiopoulos, Heileman and Huang (1991) extends the analysis of this same ART 1 subclass under fast learning conditions for patterns of different sizes. They derive a stronger upper bound for learning – self-stabilisation of an input list containing m distinct-size patterns will occur after at most m list presentations. They also show that the model creates category templates that are distinct and derives no more templates than there are unique input patterns. Thus this ART 1 subclass does not waste resources in duplicating category prototype nodes. All these properties hold independently of the order of presentation of input patterns. On the down side however it is also shown that it is possible to create templates that cannot be directly accessed by any pattern in the training set after self-stabilisation has occurred. This property is dependent upon the order of presentation of input patterns.

Georgiopoulos, Heileman and Huang (1992) further consider the issues of template creation and self-stabilisation. They disprove the general case of the N-N-N conjecture that in ART 1 "...if the F2 layer has at least N nodes, then each member of a list of N input patterns, which is cyclically presented at the F1 layer, will have direct access to an F2 layer node after at most N list presentations" (Georgiopoulos, Heileman and Huang, 1992, p.746).

Their previous work showed the validity of this conjecture on a subclass of ART 1 models possessing small values of parameter L (associated with adaptation of bottom-up LTM traces.) However, with large values for this parameter explicit counter-examples can be created that violate the N-N-N conjecture.

They conclude that with such large L values, templates can be created that are not directly accessed by training input patterns after self-stabilisation occurs. Furthermore, it is also possible for the number of templates to exceed the number of distinct input patterns after self-stabilisation. (Note that the former but not the latter conclusion was shown to be true in previous work using small L values.) These results may have negative implications for the use of ART 1 in some circumstances. If the N-N-N conjecture had proved to be generally true the upper bounds on both training cycles and number of F2 nodes required would have been greatly reduced for all ART 1 networks.

Other work considers explicit modifications to ART 1. For example, Ryan and Winter (1987) demonstrate that ART 1 does not resolve the stability-plasticity dilemma as effectively as originally claimed. Early in the training phase it is actually possible for a recognition node to be recoded so that it responds differently to the same input pattern after a sequence of intervening presentations. This occurs if the intervening inputs all match to the recognition node that coded for the original pattern, but offer gradual variations on the category that move it away from the original pattern. Ultimately this will lead to the node ceasing to include the original pattern within its recognition category.

Ryan and Winter describe an improved version of ART 1 that introduces variable thresholds into the matching between top-down and bottom-up inputs. This is called symmetric adaptive thresholding and prevents the recoding problem described above. Their network also shows a reduced number of resets during training in comparison to the original version of ART 1. Other advantages are the ability to cope with continuously presented input patterns, and short presentation intervals for each pattern.

Ryan (1988) further develops this work in a network architecture which can cope with binary or analogue input patterns. The so-called Resonance Correlation Network (RCN) remains architecturally closer to ART 1 than ART 2 however. Noise is suppressed by normalising the LTM weights during learning, rather than by complicating the input layer to normalise stimuli prior to learning.

RCN further differs from ART 1 by removing the sequential search through the category nodes following a mismatch. In such cases, an unused node from the category recognition layer is immediately selected to code for the novel input pattern. RCN also retains the ability from Ryan's and Winter's earlier system to cope with continuous presentation of input patterns.

Stork (1989) observes that ART 1 is not a panacea for pattern recognition problems. In particular, he criticises the ART 1 property of self-scaling, where novelty is determined by the orienting system based on the percentage of difference between an input pattern and a category prototype. Stork identifies a number of applications where this property causes ART 1 to fail to learn to desired categorisation.

For example, in character recognition being able to discriminate between the letters "O" and "Q" requires attending to the features comprising the diagonal bar in the "Q". However, if these same features occur over the letter "I" they should be dismissed as noise and not result in the formation of separate categories for "I" and "noisy I". A fixed vigilance level cannot simultaneously achieve these two desiderata. Low vigilance allow the "noisy I" to be correctly

categorised, but erroneously merges "O" and "Q" into the same category. Conversely, high vigilance allows "O" and "Q" to be discriminated but causes "noisy I" to acquire its own category node separate to "I". This problem can of course be overcome by external control of the vigilance parameter during training. The teacher can raise the vigilance level when "Q" occurs as an input pattern, and relax vigilance for "noisy I".

However, Stork also shows that there are other problems on which ART 1 fails even when supervised learning is allowed. One such example is the XOR problem, infamous in neural network research, which requires maximally dissimilar inputs {0,1} and {1,0} to be mapped to the same recognition category {1}.

One can however question the appropriateness of Stork's criticisms. ART 1 is at heart an unsupervised learning system. Its function, therefore, is to learn to recognise significant regularities in its input environment. Thus, it is hardly shocking that ART 1 is incapable of forming arbitrary mappings that require these similarities to be ignored in assigning input patterns to prescribed categories. In such applications ART models specifically intended for supervised learning (such as ARTMAP) should be employed.

Similar ground to Ryan and Winter (1987) and Stork (1989) is covered in Moore (1989), although in terms of pattern clustering. Here, the ART 1 learning algorithm is abstracted from its architecture in order to analyse its pattern clustering properties. This reveals the self-scaling problem, where vectors of larger magnitude can have larger cluster sizes (due to greater noise tolerance). It also shows problems with the stabilisation of code learning in ART 1. Moore proposes two forms of stability that a learning mechanism should provide. Firstly, individual clusters must be stable (i.e. category prototypes should not be re-written so that previous inputs cease to be exemplars of that category). Secondly, the total number of clusters should be stable, such that a finite number of clusters are formed even with an infinite set of inputs. ART 1 is shown to fail to achieve both these desiderata in certain circumstances. (The first form of stabilisation failure is of course the same as that demonstrated by Ryan and Winter.)

Levine (1989) considers the need for selective attention in ART models. Biasing a network to pay more attention to some features over others offers an alternative solution (that requires less direct supervision) for problems of self-scaling described above. Levine suggests two means to this end. The first is to gate all connections between the F1 and F2 layers by "attentional bias nodes" that provide variable modulation for different features. These nodes are themselves adjusted by external reinforcement and "habit nodes". External reinforcement allows a teacher to specify the correctness of a classification decision, while habit nodes record which criteria (feature types) were used in earlier category selections. An alternative modification is to bias links from F1 nodes to the reset node instead of directly to F2 nodes. This requires the F1 nodes to be divided into sub-fields based on feature type, and then different gain levels can be used for different feature types.

Levine also suggests modifications to ART 1 and ART 2 that will support ambiguity detection. These ART models always make a definite category decision, and do not take into account whether the input is a prototypical or borderline member of the category. (The latter being a potential member of another category and thus ambiguous.) To provide such ambiguity detection Levine adds an extra layer of nodes F2' in one-to-one mapping with F2, and two levels of vigilance, Gmax and Gmin. The intuition behind the latter is that if a comparison between a category prototype and an input pattern, I, passes Gmax then an unambiguous classification of I can be made as normal. Similarly if the comparison fails Gmin then I is definitely not a member of the category and further search occurs. However, if the comparison falls between Gmax and Gmin a tentative (and possibly ambiguous) classification

has occurred. The F2' node equivalent to the F2 node that was used in the match is then activated, and further search through other F2 nodes occurs. An ambiguity detector is activated in turn if two or more F2' nodes become active. When this occurs the F2 node that was most highly activated is chosen as the classification of I as usual, but the ambiguity detector suppresses learning for the LTM weights. This prevents uncertain matches causing a shift in category prototypes.

Levine and Penz (1990) further describe ART 1.5, a model that is intermediate to ART 1 and ART 2 and is used to classify radar pulse signals. ART 1.5 uses the same learning rule as ART 2 and processes real-valued data, However since the input data is not overly complex, as with ART 1, no pre-processing is performed. The model also employs the dual vigilance levels, Gmax and Gmin, described above. In this instance however they are not used to detect ambiguity but to correct for the non-uniform clustering of the input patterns used in the problem.

Dual vigilance parameters are also employed in an extension of ART 1 described in Shih, Moh and Chang (1992). This goes even further however in that each parameter is used in separate matching tests. The authors observe that the standard match used in ART 1 counts how many features an input pattern possesses that are present in the category prototype, but not vice versa. For some applications, such as character recognition, this creates problems. For example the character "F" has all the same features as "E", plus others. This means that if these characters are presented to ART 1, then successful category formation will be dependent the order of presentation of the inputs. If "E" precedes "F" in the training set, then "E" and "F" will be merged into a single category, since "F" passes the feature count for the category formed by "E". If "F" precedes "E" however, separate categories are likely to be formed, since "E" has features not present in "F". Shih et al therefore introduce bi-directional matching, using an additional vigilance test to count the features present in the category prototype that are not in the input pattern. Only if both tests are passed (within the set vigilance levels) is the input then accepted as an exemplar of that category,

7.2 Variants of ART 2

Other work has looked solely at ART 2. For instance, Burke (1991) analyses ART 2 in terms of conventional pattern clustering algorithms. She makes the observation that in fast learning mode, search following a reset due to a mismatch is redundant. The initial choice of category will be the closest match from the available prototypes. If this match is not acceptable therefore, a new category should immediately be formed without searching the other categories available. Burke suggests this can be achieved by explicitly marking F2 nodes as "committed" (i.e. coding for a particular category prototype) or "uncommitted". Resets now occur only after failed matches to committed nodes, and cause a previously uncommitted node to learn the novel input. It is also suggested that a "residue" memory be maintained to store patterns which do not match to any category once all F2 nodes are committed. This can then be used to control the vigilance parameter; if the residue becomes too large, vigilance can be lowered and the unmatched pattern re-input.

Peper, Shirazi and Noda (1993) suggest a more effective method of noise suppression than that used in ART 2-A. They observe that the measure used to judge the match between two patterns is based upon their vector angle. They present a measure that, while primarily still based upon the vector angle, also takes into account the length of the patterns. This is of importance because in some cases patterns are subject to noise regardless of their size, and in

such cases shorter patterns will be more affected by the same level of noise than longer ones. The new measure therefore ensures short patterns have less effect upon LTM weight changes.

A more radical variant of ART 2 is proposed by Fujita and Bavarian (1991). This replaces the "winner-take-all" learning rule with a topology preserving mapping (TPM) rule. Initially developed by Kohonen (c.f. Kohonen, 1989), TPM differs from winner-take-all in that "neighbouring" nodes to the maximally activated node are also allowed to adjust their LTM weights. ("Neighbouring" nodes being those which have similar feature detectors to the winner and are thus also highly activated by the input pattern.) ART 2-TPM is demonstrated against ART 2 in a difficult recognition problem that requires line orientation detectors to be formed. ART 2-TPM proves superior for this task, showing more effective self-organisation and faster convergence than ART 2.

7.3 Variants of Supervised ART Models

Other research has considered teacher-directed learning in ART. Georgiopoulos, Huang and Heileman (1994) perform a formal analysis of ARTMAP's learning properties in a similar manner to their previous work on ART 1 (Georgiopoulos, Heileman and Huang, 1990, 1991; described in section 7.1). Previously discovered properties of ART 1 are not however directly transferable to ARTMAP owing to the latter's use of variable ARTa vigilance during training. Nonetheless, Georgiopoulos et al find very similar learning properties for ART 1 and ARTMAP.

They consider the case of fast learning of a many-to-one mapping in ARTMAP and prove that this can always be achieved in a maximum of M_a-1 presentations of a list of inputs, where M_a is the number of one bits in each one of the input patterns. (The mapping is deemed to be completed when all ARTa input patterns in the list can directly access their correct equivalent ARTb F2 category, and application of the learning rule causes no actual change in the weights.) Georgiopoulos et al also prove a stronger result for the special case of $M_a\rho_a > 1$. Under this circumstance, learning can terminate after $(M_a - M_a\rho_a)$ list presentations.

Kasuba (1993) describes a simplified version of fuzzy ARTMAP. This has only one user-changeable parameter (ARTa vigilance) and fixes all other parameters to their standard default values. In addition, simplified fuzzy ARTMAP does not perform self-organisation of the teaching inputs at ARTb, instead such inputs are taken to be exact category representations. This simplified version of fuzzy ARTMAP is claimed to offer comparable computational power to the original but with greater computational efficiency.

Healy, Caudell and Smith (1993) describe LAPART (LAterally Primed ART) which performs supervised learning between two ART 1 modules akin to ARTMAP. The primary difference with ARTMAP is that LAPART has no intermediate map field. Instead, each category node in the ARTa module has direct feedforward connections to every category node in the ARTb module. ARTa is also in control of the presentation onset for stimuli to ARTb.

The behaviour of LAPART is quite similar to ARTMAP in the case where the input pattern to ARTa, Ia, is novel. A new category node is selected at ARTa, and this is associated with the category node activated at ARTb by its own input pattern, Ib. Of course, in LAPART this association is learned by adjusting the weights of the direct connections between the F2 layers of ARTa and ARTb rather than through the indirect formation of an association at the map field.

In the case of a recognised Ia, LAPART's performance differs markedly from ARTMAP. The F2 node activated at ARTa by Ia primes the associated node at ARTb while inhibiting

presentation of the input, Ib, to ARTb. Simultaneous matching of Ib with the top-down expectation that resulted from priming by ARTa then occurs. Thus, in contrast to ARTMAP, in this case ARTb does not operate autonomously in selecting an F2 node for Ib. Rather it is guided by lateral priming from ARTa. However, ARTb is still free to override the priming if too large a mismatch occurs at the input layer. If this happens resets are sent to both ART 1 modules and an entirely new association must be learned. If the match is accepted, the existing association is reinforced or refined.

The inability of LAPART's ARTb module to have direct access to its recognition codes with familiar Ia's is not necessarily a drawback. For example, the application domain described by Healy et al involves making predications about likely event sequences. The category at ARTa is meant to represent an event class occurring at time t, while the associated category at ARTb is the event class which should follow at t + 1. Priming by ARTa allows the predicated event to be the first one tested by ARTb, thus bypassing unnecessary computation in cases where the prediction proves true.

Baxter (1991) describes another supervised ART architecture. A class of networks called ECART (Error Correction ART) are introduced. These make explicit error computations between input patterns and learned templates at a separate error field. This provides the inputs to a novelty detector which can incorporate variable attention to input features. The architecture is otherwise similar to single ART module which performs unsupervised learning. A supervised ECART network is then achieved by adding an extra output layer connected to the F2 layer at which teaching inputs are presented. Every node in the F2 layer has adaptive connections to each node in the output layer. Associative learning occurs at these connections using a Hebbian learning rule (i.e. weights are increased at a connection when both nodes are simultaneously active). Baxter claims that supervised ECART is able to solve the XOR problem using a single trial.

References

Y.F. Asfour, G.A. Carpenter, S. Grossberg and G.W. Lesher (1993) Fusion ARTMAP: A Neural Network Architecture for Multi-Channel Data Fusion and Classification, Proc. World Congress on Neural Networks, vol. II, 210-215.

Baxter, R.A. (1991) Error Propagation and Supervised Learning in Adaptive Resonance Networks, IJCNN-91, Volume II, 423-429.

Beale, R. and Jackson, T. (1990) Neural Computing: An Introduction. Bristol: IOP.

Burke, L.I. (1991) Clustering Characterization of Adaptive Resonance, Neural Networks, 4(4). 485-491.

Carpenter, G.A. and Grossberg, S. (1987a) A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine, Computer Vision, Graphics and Image Processing, 37, 54-115. Reprinted in Grossberg (1988) 251-315, and Carpenter and Grossberg (1991) 316-382.

Carpenter, G.A. and Grossberg, S. (1987b) ART 2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns, Applied Optics, 26(23), 4919-4930. Reprinted in Carpenter and Grossberg (1991) 398-423.

Carpenter, G.A. and Grossberg, S. (1988) The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network, Computer, 21(3), 77-88.

Carpenter, G.A. and Grossberg, S. (1990) ART 3: Hierarchical Search Using Chemical Transmission in Self-Organizing Pattern Recognition Architectures, Neural Networks, 3(2), 129-152. Reprinted in Carpenter and Grossberg (1991) 453-499.

Carpenter, G.A. and Grossberg, S., eds (1991) Pattern Recognition by Self-Organizing Neural Networks. Cambridge, MA: MIT Press.
Carpenter, G.A. and Grossberg, S. (1992) A Self-Organizing Neural Network for Supervised Learning, Recognition and Prediction, IEEE Communications Magazine, 30(9), 38-49.
Carpenter, G.A. and Ross, W.D. (1993) ART-EMAP: A Neural Network Architecture for Learning and Prediction by Evidence Accumulation, Proceedings of the World Congress on Neural Networks, Volume III, 649-656.
Carpenter, G.A. and Tan, A.H. (1993) Rule Extraction, Fuzzy ARTMAP, and Medical Databases, Proceedings of the World Congress on Neural Networks, Volume I, 501-506.
Carpenter, G.A., Grossberg, S. and Reynolds, J.H. (1991) ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network, Neural Networks, 4(5), 565-588. Reprinted in Carpenter and Grossberg (1991) 503-544.
Carpenter, G.A., Grossberg. S. and Rosen, D.B. (1991a) ART 2-A: An Adaptive Resonance Algorithm for Rapid Category Learning and Recognition, Neural Networks, 4(4), 493-504.
Carpenter, G.A., Grossberg, S. and Rosen, D.B. (1991b) Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System, Neural Networks, 4(6), 759-771.
Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H. and Rosen, D.B. (1992) Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps, IEEE Transactions on Neural Networks, 3(5), 698-712.
Fujita, M. and Bavarian, B. (1991) An ART2-TPM Neural Network for Automatic Pattern Classification, IJCNN-91, Volume II, 479-484.
Georgiopoulos, M., Heileman, G.L. and Huang, J. (1990) Convergence Properties of Learning in ART1, Neural Computation, 2(4), 502-509.
Georgiopoulos, M., Heileman, G.L. and Huang, J. (1991) Properties of Learning Related to Pattern Diversity in ART1, Neural Networks, 4(6), 751-757.
Georgiopoulos, M., Heileman, G.L. and Huang, J. (1992) The N-N-N Conjecture in ART1, Neural Networks, 5(5), 745-753.
Georgiopoulos, M., Huang, J. and Heileman, G.L. (1994) Properties of Learning in ARTMAP, Neural Networks, 7(3), 495-506.
Grossberg, S. (1987) Competitive Learning: From Interactive Activation to Adaptive Resonance, Cognitive Science, 11(1), 23-63. Reprinted in Grossberg (1988) 213-250.
Grossberg, S., ed (1988) Neural Networks and Natural Intelligence. Cambridge, MA: MIT Press.
Healy, M.J., Caudell, T.P. and Smith, S.D.G. (1993) A Neural Architecture for Pattern Sequence Verification Through Inferencing, IEEE Transactions on Neural Networks, 4(1), 9-20.
Kasuba, T. (1993) Simplified Fuzzy ARTMAP, AI Expert, 8(11), 18-25.
Kohonen, T. (1989) Self-Organization and Associative Memory, 3rd Edition. Berlin: Springer-Verlag.
Levine, D.S. (1989) Selective Vigilance and Ambiguity Detection in Adaptive Resonance Networks, in Webster (1989) 1-7.
Levine, D.S. and Penz, P.A. (1990) ART 1.5 - A Simplified Adaptive Resonance Network for Classifying Low-Dimensional Analog Data, IJCNN-90, Volume II, 639-642.
Moore, B. (1989) ART 1 and Pattern Clustering, in Touretzky, Hinton and Sejnowski (1989) 174-185.
Mozer, M., Smolensky, P., Touretzky, D., Elman, J. and Weigend, A., eds (1994) Proceedings of the 1993 Connectionist Models Summer School. Hillsdale, NJ: Lawrence Erlbaum Associates.

Peper, F., Shirazi, M.N. and Noda, H. (1993) A Noise Suppressing Distance Measure for Competitive Learning Neural Networks, IEEE Transactions on Neural Networks, 4(1), 151-153.
Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Internal Representations by Error Propagation, in Rumelhart and McClelland (1986) 318-362.
Rumelhart, D.E. and McClelland, J.L., eds (1986) Parallel Distributed Processing - Explorations in the Microstructure of Cognition, Volume 1: Foundations. Cambridge, MA: MIT Press.
Rumelhart, D.E. and Zipser, D. (1985) Feature Discovery by Competitive Learning, Cognitive Science, 9(1), 75-112. Reprinted in Rumelhart and McClelland (1986) 151-193.
Russell, I.F. (1991) Self-Organising Competitive Learning and Adaptive Resonance Networks, International Journal of Neural Networks Research and Applications, 2(2-4), 67-73.
Ryan, T.W. (1988) The Resonance Correlation Network, Proceedings of the IEEE Second International Conference on Neural Networks, Volume I, 673-680.
Ryan, T.W. and Winter, C.L. (1987) Variations on Adaptive Resonance, Proceedings of the IEEE First International Conference on Neural Networks, Volume II, 767-775. Reprinted in Carpenter and Grossberg (1991) 385-396.
Shih, F.Y., Moh, J. and Chang, F. (1992) A New ART-Based Neural Architecture for Pattern Classification and Image Enhancement without Prior Knowledge, Pattern Recognition, 25(5), 533-542.
Stork, D.G. (1989) Self-Organization, Pattern Recognition, and Adaptive Resonance Networks, Journal of Neural Network Computing, 1(1), 26-42.
Tan, A.H. (1994) Rule Learning and Extraction with Self-Organizing Neural Networks, in Mozer, Smolensky, Touretzky, Elman and Weigend (1994), 192-199.
Touretzky, D., Hinton, G. and Sejnowski, T., eds (1989) Proceedings of the 1988 Connectionist Models Summer School. San Mateo, CA: Morgan Kaufmann.
Webster, W., ed (1989) Simulation and AI 1989. San Diego: Society for Computer Simulation.