ELSEVIER

Cognition 85 (2002) B43–B50

COGNITION

www.elsevier.com/locate/cognit

Brief article

Learning and development in neural networks – the importance of prior experience

Gerry T.M. Altmann

Department of Psychology, University of York, Heslington, York, YO10 5DD, UK

Received 6 February 2001; accepted 23 May 2002

Abstract

Infants can discriminate between familiar and unfamiliar grammatical patterns expressed in a vocabulary that is distinct from that used earlier during familiarization (*Cognition 70*(2) (1999) 109; *Science 283* (1999) 77). Various models have captured the data, although each required that discrimination be distinct, in terms of the computational process, from familiarization. This article describes a simple recurrent network (SRN), equipped only with the assumption that it should predict what comes next, which models the data without distinguishing between familiarization and discrimination. To accomplish this, the SRN requires pre-training on a range of sequences instantiating different structures and different vocabulary items to those used subsequently during familiarization and test. Pre-training enables the network to avoid replacing structure acquired during familiarization with structure experienced at test. An equivalent enabling condition may underpin infants' resistance to catastrophic interference between the different structures and vocabulary items to which they are exposed. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Grammar learning; Neural networks; Catastrophic forgetting; Infant acquisition

1. Introduction

Marcus, Vijayan, Bandi Rao, and Vishton (1999) described a study in which 7-month-old infants were familiarized with sequences of syllables generated by an artificial grammar; the infants were then able to discriminate between sequences generated both by that grammar and another, even though the sequences in the familiarization and test phases were instantiated in different sets of syllables (cf. earlier work by Gomez & Gerken, 1999). In one study, infants were familiarized either to sequences such as '*ga ti ga*' (an 'ABA' pattern) or to sequences such as '*ga ti ti*' (an 'ABB' pattern). In a subsequent test phase containing both these patterns, infants attended more to whichever sequences did not obey

E-mail address: g.altmann@psych.york.ac.uk (G.T.M. Altmann).

PII: S0010-0277(02)00106-3

the pattern to which they had previously been familiarized, despite the fact that the test sequences were instantiated in a new set of syllables (e.g. '*wo fe wo*' or '*de ko ko*'). Marcus et al. claimed that such behaviours were beyond the capabilities of certain kinds of statistical learning mechanisms. This last claim provoked various demonstrations that neural networks *could* model the infant behaviour. However, each such demonstration had its shortcomings. Thus, Christiansen and Curtin (1999) switched off learning during the test phase, but failed to establish that the information that was required to distinguish between the ABA and ABB test patterns would be maintained if learning were not switched off (cf. catastrophic forgetting – see below). Seidenberg and Elman (1999) first trained their network to explicitly recognize whether a syllable in a sequence was the same as the preceding one. Subsequently, the network was trained to categorize ABA and ABB sequences as belonging to one pattern or the other, and at test, the network correctly assigned the right pattern to each sequence. However, the training regime here was quite different to that employed in the Marcus et al. study. Gasser and Colunga (1999) explicitly coded whether one word in a sequence was the same as or different to another, whilst the Shultz (1999) and Shastri (1999) models required the explicit coding of position (e.g. 'word 1', 'word 2'). In each case assumptions had to be made in respect of the models' coding of the input (cf. also Negishi, 1999). The Altmann and Dienes (1999) model, and its shortcomings, are described below.

This article describes a new method for training a neural network which enables the modelling of the Marcus et al. (1999) data. No network parameters need be changed between familiarization and test (learning is never switched off), no special assumptions need be made regarding the coding of the input, and no explicit 'teacher' is required. The manner in which the model succeeds sheds light on various principles that underlie learning in neural networks which, if they generalize to human learning, have significant implications for the manner in which infants learn new sentence structures without unlearning old ones.

2. Background to the model

Altmann and Dienes (1999) modelled the Marcus et al. (1999) data using a simple recurrent network (SRN) (Elman, 1990). The input and output units encoded a localist representation of the eight syllables used by Marcus et al. for the familiarization phase, the eight (different) syllables they used for the test phase, and two (largely redundant) markers for the beginning and end of each sequence. These input units fed into a recoding layer (20 units) which in turn fed into a further hidden layer (25 units) – the recurrent layer – whose output proceeded both to the networks' output layer (with the same number of units as the input layer) and to a set of copy units which fed back into this layer.

The model was trained on sequences from either the ABA or ABB grammar, and after each input the network had to predict the identity of the next input. At test, the network was presented with sequences drawn from both grammars but instantiated in a different set of syllables. Its task was again to predict what the next input would be. This time, the weights on the connections leading to the recurrent layer were 'frozen'; the back-propagation algorithm could only, therefore, change weights on the connections between the input and recoding layer, and between the recurrent and output layer (cf. Dienes, Altmann, & Gao, 1999). The model better learned sequences drawn from the familiar grammar than from the other grammar.

This result is unsurprising; given that the recurrent layer encodes 'structure in time', the weights on connections to and from this layer encoded the sequential structure of whichever grammar the network had previously been trained on. The connection strengths between the input and recoding layer (and the recurrent and output layer) encoded the mapping of syllables onto this structure. By freezing the connections into the recurrent layer, the grammar that was acquired during the familiarization phase was 'fixed', and the network therefore had to learn only the mapping of the new vocabulary onto this grammar. If an input sequence obeyed the grammar, a mapping would be possible and the appropriate predictions made. If an input did not obey the grammar, the task of predicting the next element in a sequence would be hindered by the requirement to map each sequence onto an underlying internal representation which did not accord with the pattern underlying the sequence.

There were two reasons for freezing the connections into the recurrent layer: first, it forced each network to map the test sequences onto whatever internal representation of sequential structure had been formed during the familiarization phase; and second, it ensured that the novel test patterns (which had not been presented in the familiarization phase) did not cause the networks to unlearn the earlier grammar and replace it with new structure (cf. French, 1999; McCloskey & Cohen, 1989). Freezing would not be required if the network could somehow partition its internal representational space in a way that was resistant to the 'unlearning' that normally results when a previously trained network is given novel stimuli to learn. This could be achieved, in principle, by densely populating that space so that any new representations that are formed by the network would integrate with existing, previously entrenched, representations. These new representations, encoded amongst the older ones, would not be so easily changed by further novel stimuli. And populating the space would better simulate the state of the infants' knowledge when they enter the experiment (cf. McRae & Hetherington, 1993).

To test the effects of a more constrained representational space, the Altmann and Dienes (1999) simulations were repeated with two changes: first, the networks were all pre-trained on an arbitrary grammar and associated vocabulary; and second, no connection weights were frozen – the same network parameters were used during familiarization and test.

3. The simulations

Sixteen neural networks based on the architecture described above were trained on the 252 different sentences generated by the grammar and vocabulary used in Elman (1990)[1]; each was assigned a different random initialization of connection weights. Ten thousand sentences (approximately 40 repetitions of each sentence) were presented to each network six times (cf. Elman, 1990). The 29 vocabulary items were encoded across 29 input units.

[1] In order to ensure replicability, each simulation was repeated three times; each was identical except for a different random initialization of each of the 16 networks' connection weights. Unless otherwise stated, the same statistical pattern was found each time, and only the first simulation in each triplet will be reported.

Learning rate and momentum were set at 0.2 and 0.01, respectively. After this pre-training phase, the 16 networks (corresponding to the 16 infants in the Marcus et al. (1999) study) were divided into two groups; one was trained on the 48 ABA sequences used in the Marcus et al. study (comprising three repetitions of 16 different sequences), and the other was trained on the 48 ABB sequences. This familiarization phase was repeated for 50 epochs (cf. Altmann & Dienes, 1999). The ABA and ABB sequences required a vocabulary of eight items, encoded across eight further input units. In the final (test) phase, each network was given a different random ordering of 12 sequences – six ABA sequences and six ABB sequences. Each test sequence was expressed in a new set of eight vocabulary items, encoded across another set of eight input units. The number of sequences in the familiarization and test phases, as well as the number of vocabulary items making up those sequences, matched those of the Marcus et al. study. At test, the network iterated around the first test sequence 20 times; on the final iteration its prediction for what should occur after 'AB' (either B or A) was compared against what did occur in that final position; then, the network moved to the next test sequence, iterated around that 20 times, and so on (see Dienes et al., 1999, for discussion). The intention was to see how well the network would learn a test pattern as a function of which grammar it had been previously exposed to.

3.1. Results and discussion

To measure the accuracy of the networks' predictions, the product moment correlation was calculated between the output vector corresponding to the network's prediction on the last iteration, and the vector corresponding to what the network should have predicted. These correlations were averaged across the six ABA test patterns and the six ABB test patterns, and were then submitted to an analysis of variance with networks as a random variable and the factors Training (ABA or ABB) and Test (ABA or ABB). There were neither effects of Training nor Test (both $F < 1$) but there was a significant interaction between the two, with a higher correlation for sequences congruent with the previously trained grammar than for sequences that were incongruent (0.91 vs. 0.87: $F(1,14) = 15.4$, $P < 0.002$). Planned comparisons confirmed that the networks were better on ABA patterns after ABA familiarization, and ABB patterns after ABB familiarization (both $P < 0.05$).

The networks thus discriminated between the test patterns as a function of whichever grammar they had previously been trained on. But to what extent did this require pre-training on the Elman sentences? To address this, new simulations were run but without pre-training: There was no effect of Training ($F < 1$), but there was a significant effect of Test ($F(1,14) = 25.8$, $P = 0.0002$), with better learning of the ABB patterns than the ABA patterns (0.94 vs. 0.88) – adjacent repetitions are generally easier to learn in an SRN than longer distance ones. There was no interaction between Training and Test ($F < 1$). Evidently, pre-training *was* required in that first simulation.

One possible confound in that first simulation was that some of the sequences used for pre-training had an underlying ABA structure (e.g. 'boy chase boy'). To ensure that these sequences were not responsible for the result, the first simulation was repeated with a new pre-training set that did not contain such sequences. As before, there were neither main effects of training pattern ($F(1,14) = 2.44$, $P > 0.1$) nor test pattern ($F < 1$), but congruent items were better learned than incongruent items (0.91 vs. 0.88; $F(1,14) = 14.47$, $P < 0.002$). Planned comparisons confirmed the effect of congruency after both ABA and ABB familiarization (both $P < 0.05$).

4. General discussion

Pre-training on the Elman sequences enabled better learning of ABA patterns after ABA familiarization, and of ABB patterns after ABB familiarization. The vocabularies to which each network was exposed were different at each phase, suggesting that the networks had abstracted something about the underlying ABA and ABB patterns in the familiarization phase, and had used this abstract representation to aid (or hinder) learning in the final phase.[2] What remains to be explained is why pre-training should have had this effect.

One possibility is that pre-training causes the network to lay down structure in the representational space that becomes particularly well entrenched on account of the diversity of the stimuli and the amount of learning. This diversity causes the representational space to become relatively densely populated (with representations of the different dependencies that can be abstracted across the stimuli); any new structure that is learned subsequently must be encoded amongst these existing, entrenched, structures. The new structure therefore inherits the entrenchment of the previously learned structure through being represented, in part, within those previously learned structures.

An alternative is that the beneficial consequences of pre-training arise from the sigmoid activation function which converts a unit's net input into its net output. If the weights on the input connections are relatively large, changes to those weights will have relatively little effect on the net output of that unit. Consequently, for units with large net inputs, the error term that is calculated by the back-propagation algorithm will be small and the weight adjustments correspondingly so. Thus, the output activations of units with large net inputs will not change as a function of learning as much as the output activations of units with small net inputs. Pre-training may simply push the weights towards larger values, and so make the network less sensitive to weight changes and less prone, therefore, to unlearning. To explore this, a simulation was run in which, after pre-training, the weights were randomly re-assigned within the network, thereby breaking up any encoded structure (connections to the copy units were left intact). If the benefit of pre-training is due to the encoding (and entrenchment) of structure, the effect of congruency should be

[2] Marcus et al. (1999) included a third experiment using AAB and ABB patterns, because infants in the first two experiments might have distinguished between ABA and ABB patterns on the basis of reduplication in one case and a lack of reduplication in the other (but see Marcus et al., 1999, note 18). The current simulations used the ABA/ABB grammars (Marcus et al., 1999; Experiments 1 and 2) because any bias at test to better predict the final element (the only element that distinguishes between the two patterns) could only be due to an influence of the familiarization phase, or to a bias to better learn adjacent repetitions than non-adjacent ones. In the crucial simulations reported in the main text, the only bias at test was to better predict the final element of ABB patterns after ABB familiarization, and the final element of ABA patterns after ABA familiarization. More generally, a recurrent network of the kind used here can no more easily learn that an element at time 1 predicts an identical element at time 2 than it can learn that an element at time 1 predicts a different (but determinate) element at time 2 (and see Marcus et al., 1999, note 20, for discussion of neural networks' insensitivity to reduplication).

eliminated by this manipulation. In the event, the effect was indeed eliminated ($F(1, 14) = 1.2, P > 0.2$), with only an effect of Test pattern ($F(1, 14) = 9.2, P < 0.01$).[3]

One further issue concerns the nature of the pre-encoded structure – given the distinction between sequential structure on the one hand, and knowledge of a vocabulary on the other, what would be the result of pre-training the network on scrambled sequences? Vocabulary information (including frequency) would be maintained, but the constraints on sequential structure would be broken. Repeating the simulations using a pre-training set in which each sentence was scrambled revealed an interaction between training and test pattern ($F(1, 14) = 6.2, P < 0.03$), an effect of test pattern ($F(1, 14) = 16.1, P < 0.002$),[4] and no effect of training pattern ($F(1, 14) = 1.4, P > 0.2$). However, planned comparisons revealed an effect of congruency only after ABB training, not after ABA training (cf. note 3). Thus, pre-learning a vocabulary alone is only partially effective; pre-learning sequential structure in addition to that vocabulary has a more reliable effect in respect of enabling the sought-after sensitivity.

Taken together, these results suggest that only pre-training on sequentially structured input permits a reliable effect of congruency to emerge in the absence of any other biases; a quite different behaviour emerges following the break-up of the structures that are laid down during the familiarization phase, or following pre-training on input that does not contain regular sequential structure.

Finally, Dienes et al. (1999) demonstrated across a range of different grammars, albeit with freezing, that a network which is exposed to a grammar expressed in one vocabulary and is then exposed to that same grammar expressed in another vocabulary can learn to map items from one vocabulary onto corresponding items from the other; thus, an appropriately sensitive statistical learning mechanism can induce the appropriate mappings across vocabularies in these cases (see Tunney & Altmann, 2001, for discussion). Hierarchical clustering of the activation patterns across the encoding and recurrent layers suggested an equivalent behaviour here – the networks clustered like-with-like; for example, each element in the familiarization sequence ‘ga ti ga’ clustered with the corresponding element from the test sequence ‘wo fe wo’. In effect, the network ‘recognized’ the underlying equivalence between the sequences. Further research is needed to explore the generalizability of these results to more complex grammars, in learning mechanisms such as the SRN described here, and in learning mechanisms such as the infants described by Marcus et al. (1999).

5. Conclusions

Newly learned dependencies can be made relatively more resistant to unlearning if the

[3] In two of the three replications, the effect of congruency was eliminated. In the third, a partial, albeit significant effect did emerge ($F(1, 14) = 11.0, P = 0.005$) – planned comparisons revealed that although ABB patterns were better learned than ABA patterns after ABB training, ABA patterns were no better learned than ABB patterns after ABA training. Thus, pushing the weights to extremes did have a partial effect in one of the three replications, although the effects of congruency in this case were neither reliably replicable nor independent of the training patterns.

[4] The effect of test pattern did not reliably replicate ($P = 0.06$ in one replication, and $P > 0.1$ in the other).

network has already encoded a sufficient number and variety of other dependencies. A similar result has been observed for simple two-layer feed-forward networks (McRae & Hetherington, 1993), although the present observation is significant in that such networks neither encode structure in time nor do they encode abstract knowledge across connections between hidden layers.

The observation that newly acquired knowledge can be more stable if a network has already learned some other body of knowledge has implications for the manner in which human learning proceeds. When acquiring a grammar, infants are exposed to a wide range of novel sentences using novel vocabulary items. Somehow, hearing new sentences and new words does not cause the infant to catastrophically unlearn syntactic dependencies he or she has learned previously. Perhaps this is because the infant already has a core body of knowledge (not necessarily language specific) which enables novel mental representations to become relatively stable in the face of novel vocabulary items and novel structures. Without a sufficient body of pre-existing knowledge, newly learned structures, and newly learned mappings between novel words and those structures, might cause earlier structures and earlier mappings to be unlearned. And just as ‘starting dense’ – acquiring novel structure within a representational system that is already densely populated – has proved critical in enabling successful simulation of data on infant grammar learning, perhaps an equivalent principle underpins infants’ resistance to catastrophic interference between the different structures to which they are exposed. The present research adds to the body of evidence which suggests that models of statistical learning can provide insights into the nature of the conditions that enable certain kinds of learning (cf. Elman, 1993; Elman et al., 1996).

Acknowledgements

Thanks to Gareth Gaskell for many useful suggestions and the loan of ‘C-Programming’, to Ken McRae for pointing me, after the event, in the direction of his earlier work, and to Apple Computers Inc. for releasing OS X. The simulations were run using a debugged version of XTlearn in batch mode, compiled within the UNIX layer of OS X. This version of XTlearn is available from the author at http://www.york.ac.uk/res/prg/networks. The work was supported by the Economic and Social Research Council (Ref. R000238380).

References

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science*, *284*, 875.

Christiansen, M. H., & Curtin, S. L. (1999). The power of statistical learning: no need for algebraic rules. *Proceedings of the 21st annual conference of the Cognitive Science Society* (pp. 114–119). Mahwah, NJ: Erlbaum.

Dienes, Z., Altmann, G. T. M., & Gao, S. -J. (1999). Mapping across domains without feedback: a neural network model of transfer of implicit knowledge. *Cognitive Science*, *23* (1), 53–82.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48* (1), 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA: MIT Press/Bradford Books.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3b*, 128–135.

Gasser, M., & Colunga, E. (1999). Babies, variables, and connectionist models. In M. Hahn & S. C. Stone (Eds.), *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (p. 794). Mahwah, NJ: Erlbaum.

Gomez, R. L., & Gerken, L. A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, *70* (2), 109–136.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80.

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. *The Psychology of Learning and Motivation*, *24*, 109–165.

McRae, K., & Hetherington, P. A. (1993). *Catastrophic interference is eliminated in pretrained networks*. Poster presented at the fifteenth annual conference of the Cognitive Science Society, Boulder, CO.

Negishi, M. (1999). Do infants learn grammar with algebra or statistics? *Science*, *284*, 435.

Seidenberg, M., & Elman, J. L. (1999). Do infants learn grammar with algebra or statistics? (letter). *Science*, *284*, 434–435.

Shastri, L. (1999). Infants learning algebraic rules. *Science*, *285*, 1673–1674.

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks. In M. Hahn & S. C. Stone (Eds.), *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp. 665–670). Mahwah, NJ: Erlbaum.

Tunney, R., & Altmann, G. T. M. (2001). Two modes of transfer in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27* (3), 614–639.