Two Apparent “Counterexamples” To Marcus: A Closer Look

Marius Vilcu (mvilcu@cs.sfu.ca)
School of Computing Science, Simon Fraser University
Burnaby, Canada, V5A 1S6

Robert F. Hadley (hadley@cs.sfu.ca)
School of Computing Science, Simon Fraser University
Burnaby, Canada, V5A 1S6

Abstract

Marcus, Vijayan, Bandi Rao, Vishton’s experiment (1999) concerning infant ability to discriminate between simple syntactic structures has prompted many connectionists to strive to demonstrate that certain types of neural networks can replicate those results. In this paper we take a closer look at two such attempts: Shultz & Bale (2001) and Altmann & Dienes (1999). We were not only interested in how well these two models matched the infants’ reported results, but also whether or not they were able to learn the grammars involved in this process. After performing an extensive set of experiments, we found that, at first blush, Shultz & Bale’s model replicated the infant’s known data, but the model largely failed to learn the grammars. We also discovered serious problems with Altmann & Dienes’ model, which failed to match most of the infant’s results and to learn the syntactic structure of the input patterns.

Introduction

The widely known Marcus et al.’s experiment on infants (1999), involving their ability to differentiate between simple grammars, has prompted not a few connectionists to demonstrate that a variety of neural networks are indeed capable of matching the infants’ behavior.

Marcus et al. (1999) familiarized 7-month-old infants with sequences of syllables (“sentences”) generated by one of two grammars: ABA or ABB, and, in another experiment, ABB or AAB (for example, “la ti ti”, “ga ga ti”, etc). During the test phase, infants were presented with novel sequences of syllables generated by both grammars, and showed an attentional preference for sentences that were constructed with the unfamiliar grammar. Marcus et al. (1999) argued that the only explanation for such behavior is that infants possess a rule-learning mechanism that is not available to connectionist models.

There have been repeated attempts to prove that neural networks are capable of doing the same kind of discrimination as infants. Among those attempts is Elman’s (1999), who first pre-trained a simple recurrent network (SRN) to distinguish whether a given syllable is identical to a previous syllable. He then trained the same network to simultaneously discriminate between sequences of syllables generated by two simple grammars (ABA and ABB). Elman claimed that his experiment showed that SRNs successfully matched the infants’ results. In a recent paper (Vilcu & Hadley, 2001), after performing numerous experiments on Elman’s model (1999), we showed that Elman’s claim was premature, and his networks performed erratically. We emphasize, however, that Elman’s model had a more difficult task to solve: unlike infants, it was trained to recognize two specific grammars at the same time. Therefore, the fact that Elman’s experiment was questionable does not mean that other connectionist models cannot effectively match the infants’ results.

In this paper we focus on other two such attempts: Shultz & Bale (2001) and Altmann & Dienes (1999). Shultz & Bale’s experiment was performed on a cascade-correlation network, and they claim that their results “show that an unstructured neural network model without symbolic rules can simulate infant familiarization and novelty results” (2001). They also argue that the network exhibits “extrapolative generalization outside the range of the training patterns” (2001). After performing an extensive set of experiments on their model, we came to the conclusion that Shultz and Bale’s claims (2001) are substantially over-stated. We found that even though this model closely mirrors Marcus et al.’s reported data (1999), it has limited generalization capabilities. As we demonstrate below, the network fails to generalize both outside of the training space (extrapolation), and within the range of the training patterns (interpolation). Granted, Shultz & Bale never explicitly claim their model learns a grammar. However, in saying that the network was able to “recognize a syntactic pattern” (2001), and had the “ability to learn multiple syntactic forms simultaneously” (2001), Shultz & Bale imply that their model learns the underlying syntactic structure of the input patterns and is able to successfully apply this knowledge to novel items. We believe that Shultz & Bale’s network (2001) behaves more like a typical pattern recognizer, whose performance is conditioned by *familiar* “shapes” (numerical contours), than a model capable of discovering *abstract* grammatical relationships. We found that, in general, test sentences closest (in Euclidian space) to the training vectors will generate the smaller network error, regardless of whether those test sentences had been generated with the familiar or unfamiliar grammar. Therefore, it is Euclidian closeness to the training data, rather than the learning of underlying structure of input patterns, which dictates the behavior of this network.

Altmann & Dienes (1999) used a modified simple recurrent network, adding a new layer of units between the input and hidden layers of nodes. This new layer encodes two different, non-overlapping domains: the training and test sets. The common encoding of the two domains facilitates the network's generalization to the test patterns. Altmann & Dienes (1999) reported good results for their simulation. They claimed they found "significantly higher correlation for congruent sequences than for incongruent ones (…), and a significantly smaller Euclidian distance between prediction and target for congruent targets than for incongruent ones" (1999). We performed numerous experiments on this model and discovered serious problems. We found that when the networks were trained with the ABB grammar, the Euclidian distance between the actual and target vectors was consistently higher for familiar sequences than for unfamiliar ones. We believe these findings are incompatible with the Altmann & Dienes' assertion that "like the infants (…), our networks successfully discriminated between the test stimuli" (1999).

Shultz & Bale's Experiment

Shultz & Bale's simulation (2001) employs an encoder version of the cascade-correlation learning algorithm. The cascade-correlation (Fahlman & Lebiere, 1990) is a generative algorithm in feed-forward networks. It creates the network topology as it learns, by adding new hidden layer units as necessary, in order to minimize the network error for the task at hand. Each new unit is installed on a separate hidden layer alone, and receives data from both the input layer and the existing hidden layers. The hidden unit that gets added to the network is chosen from a pool of candidates: the candidate unit whose activations correlate most highly with the network current error gets added to the structure.

The "encoder" version of the network precludes direct input-output connections, in order to avoid generating networks simply having connections of weight 1 between the input and output layers (see Figure 1).

This network is similar to a prior model by Shultz (1999). The only difference between the two simulations is the input representation. In the more recent experiment, Shultz & Bale (2001) used a sonority scale, each phoneme being assigned a number between –6.0 and 6.0. They claim that this encoding scheme is more "realistic" (2001) than the one used in the previous paper (Shultz, 1999), where each syllable (collection of two phonemes) was assigned a number between 1 and 8. The choice of the new encoding was based on the fact that sonority represents the "quality of vowel likeness" (2001), i.e., some phonemes can be considered to be "more vowel-like" than others. The sonority scale ranges from "low vowels", such as /a/ and /æ/ that were assigned a sonority of +6.0, to "voiceless stops", such as /p/, /t/, and /k/ that were assigned a sonority of –6.0. For example, *ga* = -5.0 6.0, *wo* = -1.0 5.0, *ti* = -6.0 4.0. A sentence consists of three such syllables, generated using one simple grammar (ABA, ABB, AAB). For example, *ga ti ga* = -5.0 6.0 -6.0 4.0 -5.0 6.0, *li na na* = -1.0 4.0 -2.0 6.0 -2.0 6.0.

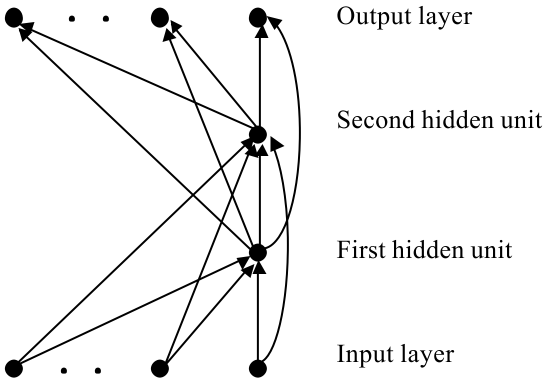

Figure 1: The cascade architecture (with 2 hidden units)

Similar to Marcus et al.'s experiment (1999), Shultz & Bales' simulation (2001) consisted of three experiments, each experiment involving 16 separate networks (one network corresponds to one infant). The first two experiments consisted of training eight networks with sentences generated by the ABA grammar, and other eight networks were trained on the ABB grammar. All 16 networks were then tested with novel sentences corresponding to both grammars. Experiment 3 was similar to the first two simulations, except that the grammars involved were AAB and ABB.

In our simulation of this model we used the same parameters as Shultz & Bale (2001): a score-threshold of 0.8, input-patience and output-patience of 1. All other training parameters were identical to Fahlman & Labiere's default values [5]. The only difference from Shultz & Bale (2001) was that we ran all experiments on double the number of networks. This permits a more accurate picture of the performance of the model.

Initially, we performed the same experiments as Shultz & Bale (2001), using identical input patterns. Each experiment consisted of training 32 networks, 16 of them with sentences generated by the ABA grammar (in the case of experiments 1 and 2) or the AAB grammar (for experiment 3), and the other 16 were trained with patterns generated by the ABB grammar (in all three experiments). In each experiment, all 32 networks were then tested with novel sentences created using both the training, familiar grammar (consistent patterns), and the unfamiliar grammar (inconsistent). Our results closely resembled Shultz & Bale's (2001), as it is shown in Table 1. The table displays the mean network error over all 32 networks for each experiment.

Table 1: Mean network error in the first three experiments

Experiment	Test vs. training sentences	Shultz & Bale's results (2001)	Our results
1 ABA vs. ABB	Consistent	8.2	8.32
	Inconsistent	14.5	16.12
2 ABA vs. ABB	Consistent	13.1	13.92
	Inconsistent	15.8	14.33
3 AAB vs. ABB	Consistent	12.9	13.63
	Inconsistent	15.3	15.05

Given the mapping between the syllables employed by Marcus et al. (1999) and the sonority values used here, and assuming that the network error resembles the time spent by infants looking at the consistent/inconsistent stimuli, it does appear that this model is capable of capturing the infants' reported data. But Shultz & Bale (2001) also claim that the network was able to generalize to novel sequences of syllables by learning the underlying structure of input patterns. As we show below, this is not the case.

The problems appeared when we altered the test patterns for each of the three experiments. When we tested interpolation (generalization within the training space), we discovered that changing just one letter in the test set does make a significant difference in the distribution of the network error.

In experiment 1, we replaced one instance of the letter "w" (sonority of –1.0) with "v" (sonority of –3.0) in one of the ABB sentences: the test sentence *wo fe fe* (-1 5 -4 5 -4 5) became *vo fe fe* (-3 5 -4 5 -4 5). If the model learned the underlying structure of the input patterns, and if it was trained on ABA patterns, i should not have any difficulty distinguishing between the unchanged ABA test sentences and the novel ABB sentences that contained one new consonant. Since this new letter was never presented to the network before, the network error was expected to be higher for the ABB sentences that contained it, along with an even better differentiation between the more familiar ABA sentences and the unfamiliar ABB sentences. In reality, our results showed that the network error for the unfamiliar ABB patterns was smaller than for ABA patterns (see Table 2).

Table 2: Network error in the three experiments using altered test patterns

Experiment	Test vs. training sentences	Altered test patterns	
		Inside the training space	Outside the training space
1	Consistent	8.83	97.29
	Inconsistent	8.46	89.83
2	Consistent	14.56	136.89
	Inconsistent	13.83	122.83
3	Consistent	14.71	144.82
	Inconsistent	14.57	129.09

In experiments 2 and 3, we replaced two instances of the letter "b" (sonority –5.0) with the letter "m" (sonority –2.0) in one ABA sentence (experiment 2), and in one AAB sentence (experiment 3): the test sentence *ba po ba* (-5 6 -6 5 -5 6) became *ma po ma* (-2 6 -6 5 -2 6) in experiment 2, and *ba ba po* (-5 6 -5 6 -6 5) became *ma ma po* (-2 6 -2 6 -6 5) in experiment 3. We do not know what infants would have done if they were presented with these new test sentences. However, to an adult the changes seem minimal, and it is entirely credible that infants would still have been able to differentiate between the familiar and unfamiliar sentences. As shown in Table 2, the model was not able to distinguish the two categories of sentences, even though the new test sentences were well within the training space. Contrary to their claim, our results showed that Shultz and Bale's model (2001) was *not* able to "generalize (…) to novel sentences within the range of the training patterns" (2001).

In order to test the extrapolation ability of the model, we re-ran the three experiments on a new test set. We picked 6 different values outside the sonority scale: 4 of these values were below -6.0 and were used as consonants: -7.0, -8.0, -9.0, -10.0, and the other 2 were greater than +6.0 and were used as vowels: 7.0 and 8.0. For instance, an ABA sentence was -10.0 8.0 -9.0 8.0 -10.0 8.0, and an ABB sentence -10.0 8.0 -9.0 8.0 -9.0 8.0. As shown in Table 2, the network error for all of the three experiments was *smaller in the unfamiliar case*. This means that the network's ability to generalize outside the training space is weak, and that the model did not reliably "recognize syntactic differences in sentences containing words with sonorities outside of the training range" (2001).

The model's inability to learn the syntactic structure of the input patterns was also shown in another experiment. We re-ran experiment 1 using two slightly more complex grammars: ABCA vs. ABCB. We used the same set of letters as in the original first experiment, but the input patterns were generated with the ABCA and ABCB grammars. After 16 different runs, the mean network error for *unfamiliar sentences was smaller* than for the familiar sentences (30.22 vs. 30.73), which means that the model did not learn the grammars.

Finally, in another set of experiments we discovered that if the network is "abstracting functions relating inputs to outputs" (2001) as Shultz & Bale claim, this kind of "abstraction" amounts to recognizing spatial shapes that are similar to the input set, rather than understanding "two syntactic forms simultaneously" (2001). In other words, we found that this network is essentially a typical shape-pattern recognition model, and not a system capable of learning grammars.

To clearly show this pattern recognition behavior, we performed two experiments: one experiment used the sonority scale encoding, and the other one used the coding scheme that Shultz made use of during a previous simulation (Shultz, 1999). That earlier model matched the newer model (Shultz & Bale, 2001) in every aspect of network structure and learning algorithm. The only difference was the input representation. Shultz (1999) assigned an odd number between 1 and 7 to category "A" syllables, and an even number between 2 and 8 to category "B" syllables. For example, *ga ti ga* was represented by *1 2 1*.

In both experiments we trained the networks on ABA generated sentences, but all training patterns had one additional property. In one experiment, the encoding value of the B syllables was always greater than the value of the A syllables (for example, 1 2 1, 3 6 3, 4 6 4, etc). In the other experiment (using the sonority scale encoding), the absolute values of both consonants and vowels were greater for the B syllables than for the A syllables (for example, -2.0 2.0 -5.0 5.0 -2.0 2.0, -1.0 1.0 -4.0 4.0 -1.0 1.0). For testing, we randomly picked 4 values within the training set (between 1 and 8 in the first experiment, and between -6.0 and 6.0 in the second one), and generated two test sets for both experiments. Each test set contained two ABA and two ABB sentences. The ABB sentences were the same in both test sets. The ABA sentences in the first test set had the same numerical contour as the training patterns (low-high-low, "peaks"), while in the second test set the ABA sentences had an opposite contour (high-low-high, "valleys") (see Table 3). We ran the two experiments on 16 different networks, and as shown in Table 3, the networks behaved differently when tested with the two sets. When the ABA test vectors represented "peaks", the network error was smaller for the familiar patterns than for the unfamiliar patterns. However, when the ABA test vectors represented "valleys", even though they had a "familiar" (ABA) structure, the error was smaller for the unfamiliar (ABB) patterns. This means that the grammatical structure does not play a significant role in the behavior of the model.

We believe these experiments demonstrate that the model was not able to develop the kind of internal representations that would enable it to actually learn the syntactic structure of the input patterns. Clearly, it is the numerical contour (shape) of the sentence that dictates the behavior of the network, and not the grammatical structure of the whole sentence. The network is an example of a traditional pattern recognition system, rather than a grammar-learning model.

Table 3: Mean network error when testing with "peaks" and "valleys" using two different input representations

Experiment	Input representation	Testing patterns		Network error
Testing "peaks"	Shultz & Bale (2001)	Consistent	-5.0 5.0 -6.0 6.0 -5.0 5.0	13.36
			-4.0 4.0 -5.0 5.0 -4.0 4.0	
		Inconsistent	-6.0 6.0 -5.0 5.0 -5.0 5.0	47.07
			-5.0 5.0 -4.0 4.0 -4.0 4.0	
	Shultz (1999)	Consistent	6 7 6	1.21
			4 5 4	
		Inconsistent	7 6 6	3.29
			5 4 4	
Testing "valleys"	Shultz & Bale (2001)	Consistent	-6.0 6.0 -5.0 5.0 -6.0 6.0	55.64
			-5.0 5.0 -4.0 4.0 -5.0 5.0	
		Inconsistent	-6.0 6.0 -5.0 5.0 -5.0 5.0	47.07
			-5.0 5.0 -4.0 4.0 -4.0 4.0	
	Shultz (1999)	Consistent	7 6 7	4.55
			5 4 5	
		Inconsistent	7 6 6	3.29

Altmann & Dienes' Experiment

Altmann & Dienes' (1999) work on simulating Marcus et al.'s experiment (1999) was based on a previous model of their own: Dienes, Altmann & Gao (1999). That previous model represented a version of a simple recurrent network that can "transfer its knowledge of artificial grammars across domains" (1999). Later, Altmann & Dienes (1999) adapted it to simulate Marcus et al.'s experiment on infants (1999). Since our intention was to analyze the models that were specifically intended to replicate the infants' results, we focused on the more recent work of Altmann & Dienes (1999).

Altmann & Dienes (1999) used a simple recurrent network with an additional layer of units between the input and hidden layers of the SRN. This additional layer is used to re-encode the input representations of two domains (the training and test domains). The function of this extra layer is to provide an abstract, common encoding of two different input sets (see Figure 2).

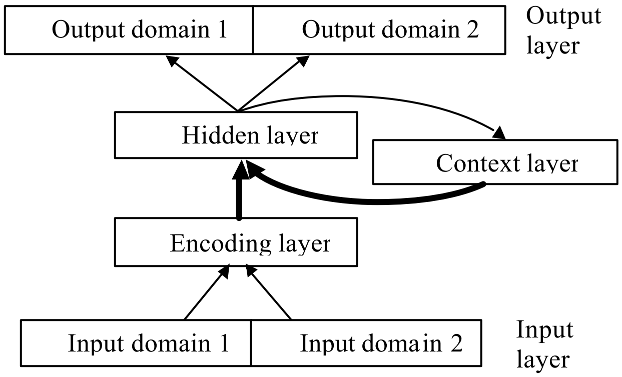

Figure 2: Version of SRN with an extra encoding layer

The connection weights between the encoding and hidden layers, as well as between the context and hidden layers, represent the "core weights", and they are frozen after training. All other connection weights represent the "mapping weights", and they are allowed to change even during testing, while the test set is learned.

Training is performed using back-propagation. The input vectors are completely orthogonal: just one input unit is active at any time, corresponding to a given syllable. Each sentence is presented to the network one syllable at a time, beginning with the activation of a special "start" unit and concluding with the activation of the "end" unit.

During "testing" on a new input set, the "core weights" were frozen, and only the "mapping weights" changed for a number of iterations, until the network learned the encoding of the new domain. After this additional learning process, all the connection weights were frozen, and the network was tested on the second domain. Although this training/test procedure may seem biologically implausible, Dienes, Almann, and Gao (1999) argue it mimics an adaptive learning mechanism, where the learning rate gradually decreases while the learning progresses. We agree that certain aspects of the adaptive learning technique may be biologically plausible, but doubt that Dienes, Altmann, and Gao's method (1999) of updating certain connection weights while keeping others frozen mirrors the human brain's activity during learning of novel patterns.

In any case, *using the same input representations and learning parameters*[1], we tried to replicate Altmann & Dienes' results (1999). We trained 16 networks on patterns generated by an ABA grammar, and another 16 networks on patterns following an ABB grammar. We then tested the networks on novel sentences having both ABA and ABB structures. It emerged that the Euclidian distance between target and prediction was always higher for patterns having the ABB structure, regardless of what the training grammar was: after ABA training, the network error for consistent/inconsistent test patters was 0.84/ 0.85, while after ABB training, the error for consistent/inconsistent test patterns was 0.87/ 0.85. Therefore, we were not able to replicate Altmann & Dienes' s results (1999). Not only was the network error for consistent test patterns very close to the error for inconsistent patterns, but also the error was higher for familiar sentences when the network was trained on ABB patterns. We also tried various other learning parameters (learning rate, momentum, number of iterations), but in each case our results showed that the model was not able to mirror infants' performance, and it is clear it did not learn the syntactic structure of the input patterns.

In passing, we note that Altmann (2002) employed a variation on the Altmann & Dienes' experimental design (1999). He pre-trained 16 different networks on simple sentences of the form Noun Verb, or Noun Verb Noun. Also, during testing he did not freeze the core weights (all the connection weights changed freely during both training and testing). After pre-training, some of the 16 networks were trained on the ABA grammar, and others were trained on sentences created with the ABB grammar. Altmann found that the model predicted the familiar test sentences better than the unfamiliar ones, concluding that once the model learns a pre-training structure, it is less likely that the test structure will replace the grammar that is learned during habituation. However, the true explanation for this behavior might stem from the fact that during pre-training the model was presented with patterns

[1] Learning rate: 0.5, momentum: 0.01, 10 iterations around each test pattern, 14 input units: 8 of them corresponded to 8 syllables of the first domain, 4 represented the four syllables of the second (testing) domain, and 2 were used to signal the start and stop of sentences. 10 hidden units were used.

resembling both ABA and ABB grammars (for instance, three consecutive sentences of the form Noun1 Verb1, Noun1 Verb2 Noun2, Noun2 Verb3 resemble an ABA pattern followed by an ABB pattern: Noun1 Verb1 Noun1, Verb2 Noun2 Noun2), and the network became attuned to both grammars. Later, during the second habituation phase, the model became increasingly biased towards the most recently trained grammar. This grammatical bias would then explain why it predicted the (most recently) familiar patterns better than unfamiliar ones.

Conclusion

In the foregoing, we have examined two connectionist models that were specifically designed to simulate Marcus et al.'s experiment on infants (1999). We considered to what degree these models were able to actually learn the grammars involved. Although we believe that the results reported by Marcus et al. (1999) do not necessarily lead to the conclusion that the infants learned the grammars[2], we wanted to see whether those connectionist models were capable of doing more than just replicate infants' results and to generalize both within and outside their training sets.

In contrast with Altmann & Dienes' model (1999), we found Shultz & Bale's (2001) very close to matching infants' performances in the Marcus et al.'s study (1999). However, the model fell short of learning the syntactic structure of the input patterns. We performed numerous experiments on their model, using various input patterns and grammars. Our results showed that their network was only driven by the numerical contours of the training patterns, and not by the generality of grammatical structure. Therefore, we conclude that Shultz & Bale's model (2001) behaves like a shape recognition system, and not like a robust model that is capable of learning grammars.

Regarding Altmann & Dienes' model (1999) we were not able to replicate their results, and we believe their model lacks consistency and robustness. We ran many experiments on their network, using various learning parameters, but we could not even mirror the infants' results. Because of this, we believe this model does not learn the syntactic structure of input patterns, and it is unable to generalize to novel items.

References

Altmann, G. T. M. (2002). Learning and development in neural networks – the importance of prior experience, *Cognition*, 85 (2), 43-50

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

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

Elman, J. L. (1999). Generalization, rules, and neural networks: a simulation of Marcus et al. , http://www.crl.ucsd.edu/~elman/Papers/MVRVsim.html

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture, *Advances in Neural Information Processing Systems*, 2, Los Altos, CA: Morgan Kaufmann, 524-532

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

Shultz, T. R. (1999). Rule learning by habituation can be simulated in neural networks, *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 665-670), Mahwah NJ: Lawrence Erlbaum

Shultz, T. R., & Bale, A. C. (2001). Neural network simulation of infant familiarization to artificial sentences: rule-like behavior without explicit rules and variables, *Infancy*, 2, 501-536

Vilcu, M., & Hadley, R. F. (2001). Generalization in simple recurrent networks, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 1072-1077), Mahwah NJ: Lawrence Erlbaum

[2] Some connectionists argue that infants could just detect whether the last two syllables in a sentence are identical, and they do not necessarily implement a rule.