Generative Connectionist Models of Cognitive Development: Why They Work

Thomas R. Shultz **David Buckingham** **Yuriko Oshima-Takane**

LNSC, Department of Psychology
McGill University
1205 Penfield Ave.
Montreal, Quebec, Canada H3A 1B1
shultz@psych.mcgill.ca
dave@hebb.psych.mcgill.ca
yuriko@hebb.psych.mcgill.ca

Abstract

Although the reasons for the success of computer simulations of psychology are often difficult to identify, it is possible to make some progress through systematic experimentation. Reasons for the success of cascade-correlation models of cognitive development are identified in two case studies. Cascade-correlation is a generative neural network model that constructs its own topology as it recruits hidden units. Capturing correct stage sequences in the integration of velocity, time, and distance cues requires a system that grows in computational power while it learns. Static networks are either too weak or too powerful to capture the full range of stages. Simulating the variation and stages in acquisition of the semantics of English personal pronouns requires sensitivity to differing amounts of addressed and non-addressed speech. Just as with children, networks benefit from the opportunity to hear personal pronouns used in exchanges between o ther speakers. Other simulations suggest that it is important for neural networks to be able to abstract regularities from the environment in order to achieve rulelike behavior and to compute unit activations in a continuous manner to simulate perceptual effects.

Why Models Work

This paper is directed towards the issue of why a particular class of computational models captures phenomena in human cognitive development. With a number of colleagues, we have applied a generative connectionist algorithm called cascade-correlation to several domains in cognitive development (Shultz et al. 1995).

The issue of why a model works is complicated. The success of a complex model of several phenomena may depend on multiple factors, some domain-general and others domain-specific. The paper begins with an overview of the cascade-correlation algorithm. Then it examines some of the reasons for the success of cascade-correlation models in two developmental domains: integration of velocity, time, and distance cues, and pronoun acquisition. In these cases, it is possible to identify some of the reasons for the success of models through systematic experimentation, in which key features of the models are varied across different conditions. This is followed by a more general characterization of cognitive developmental phenomena captured by cascade-correlation networks and the likely reasons for their capture.

The Cascade-correlation Algorithm

Like other generative neural network algorithms, cascade-correlation builds its own network topology by recruiting new hidden units as it needs them to solve a problem (Fahlman & Lebiere 1990). It starts with a minimal network of input units and output units. During training, the algorithm adds hidden units one-by-one, installing each new hidden unit on a new layer of the network. From a developmental point of view, the importance of generative connectionist algorithms like cascade-correlation is that they are able to simulate underlying developmental changes that are either qualitative or quantitative.

There are two alternating, recurrent phases in cascade-correlation learning: an output phase in which connection weights entering output units are adjusted in order to reduce the network's error, and an input phase in which new hidden units are selected and installed in the network. During the output phase, connection weights going into output units are adjusted according to a gradient descent procedure known as quickprop (Fahlman 1988). Quickprop modifies each connection weight to lower the error at the network's output units. Error is computed as the sum of squared differences between the output activations the network should be producing and those it is actually producing. Both first and second derivative information from the error function are used to compute connection weight changes. Weight changes are proportional to the negative of the slope and inversely proportional to the estimated curvature of the error function. This allows connection weight changes to be decisive and effective.

When error is no longer decreasing or the problem has not been solved in some specified number of passes through the training examples (epochs), there is a shift to the input phase. In the input phase, a pool of candidate hidden units receives trainable input from the input units and any existing hidden units. The candidate hidden units are not yet connected to the output units. The purpose of the input phase is to recruit a hidden unit whose activations correlate highly with errors at the output units. Connection weights into the candidate units are adjusted using quickprop to increase correlations between activations on the candidate units and the network's current error. When the correlations are no longer increasing or a set number of epochs has occurred, the candidate hidden unit whose activations have come to correlate best with the network's current error is selected for installation. Selected hidden units are installed into the network in a cascade, such that each new hidden unit receives input from the input units and from any previous hidden units. After installation of a new hidden unit, the algorithm reverts back to the output phase.

Thus, cascade-correlation searches not only connection weight space, but also the space of network topologies. The algorithm efficiently finds a network topology and a set of connection weights to solve the problem it is being trained on.

Although a variety of unit activation functions are available in cascade-correlation, our simulations typically use sigmoid activation functions for both hidden and output units. Occasionally, when the task is to predict quantitative output values, we use linear activation functions for output units.

Now we examine two case studies of cascade-correlation simulations in some detail, with the aim of explaining key features of their success.

Integration of Velocity, Time, and Distance Cues

In classical physics, velocity is defined as the ratio of distance traveled to the time of the journey: velocity = distance / time. Thus, distance = velocity x time, and time = distance / velocity. Some of the best evidence on children's acquisition of these relations was collected by Wilkening (1981), who asked children to predict one dimension (e.g., time) from knowledge of the other two (e.g., velocity and distance). For example, three levels of velocity information were represented by the locomotion of a turtle, a guinea pig, and a cat. These three animals were described as fleeing from a barking dog, and the child was asked to imagine these animals running while the dog barked. The child's task was to infer how far an animal would run given the length of time the dog barked, an example of inferring distance from velocity and time cues.

Cascade-correlation networks learning similar tasks typically progress through an identity stage (e.g., velocity = distance), followed by an additive stage (e.g., velocity = distance - time), and finally the correct multiplicative stage (e.g., velocity = distance / time) (Buckingham & Shultz 1994). Many of these stages have been found with children (Wilkening 1981 1982), and others remain as predictions for future psychological research. As with the children in Wilkening's (1981) experiment, the networks learned to predict the value of one dimension from knowledge of values on the other two dimensions.

Figure 1 shows rule diagnosis in a representative cascade-correlation network learning all three inference tasks. Rule diagnosis is based on correlations between network outputs and various algebraic rules like those observed in children, calculated every fifth epoch during training. To characterize network performance, an algebraic rule had to correlate positively with network responses, account for more than one-half of the variance in network responses, and account for more variance than any other rules did. For velocity and time inferences, this network exhibited an identity rule, followed by a difference rule, followed in turn by the correct ratio rule. Results were similar for distance inferences, except that there was no identity rule. There is no reason a network should favor either velocity or time information in making distance inferences because both velocity and time vary proportionally with distance. Virtually all of the cascade-correlation networks we ran showed these orderly stage progressions.

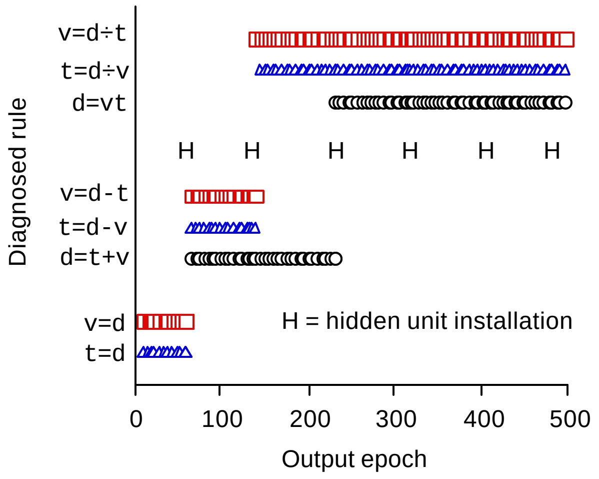

Figure 1. Diagnosis of velocity, time, and distance rules in a cascade-correlation network.

Such rule progressions are natural for cascade-correlation networks. The shift from linear to nonlinear solutions occurs because of the continued recruitment of hidden units. Linear rules include identity (e.g., time = distance), sum (e.g., distance = velocity + time), and difference (e.g., velocity = distance - time) rules, whereas nonlinear rules include product (e.g., distance = velocity * time) and ratio (e.g., time = distance / velocity) rules.

Because the sum and difference rules of the second stage are each linear, it is difficult to see at first glance why they require a hidden unit. The reason is that networks without a hidden unit are unable to simultaneously encode the relations among the three dimensions for all three inference types. In distance inferences, distance varies directly with both velocity and time. However, in velocity inferences, distance and time vary inversely; and, in time inferences, distance and velocity vary inversely. Networks without hidden units are unable to encode these different relations without a hidden unit. The first hidden unit that is recruited differentiates distance information from velocity and time information by learning weights with one sign from the former input and opposite signs from the latter inputs. This enables the network to consolidate the different directions of relations across the different inference types.

In contrast to generative networks, static back-propagation networks seem unable to capture these stage sequences (Buckingham & Shultz 1995). If a static back-propagation network has too few hidden units, it fails to reach the correct multiplicative rules. An example is shown in Figure 2 of a network with one hidden unit that fails to reach any multiplicative stages and most additive stages. If a static back-propagation network has too many hidden units, it fails to capture the intermediate additive stages on velocity and time inferences. Figure 3 shows an example of a network with two hidden units that fails to capture additive velocity and time stages.

Extensive exploration of a variety of network topologies and variation in critical learning parameters led us to conclude that there seems to be no pre-designed static back-propagation network topology that can capture all three types of stages on these tasks. Even the use of cross-connections that bypass hidden layers, another standard feature of cascade-correlation, failed to improve the stage performance of back-propagation networks. If one hidden unit provides too little power, and two hidden units provide too much power, there is no number of hidden units capable of providing the right amount of computational power. Thus, it can be concluded that the ability to grow in computational power is essential in simulating stages in the integration of velocity, time, and distance cues.

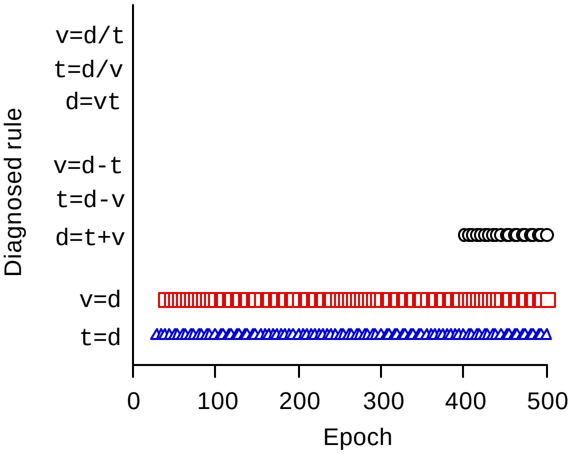

Figure 2. Diagnosis of velocity, time, and distance rules in a static network with one hidden unit.

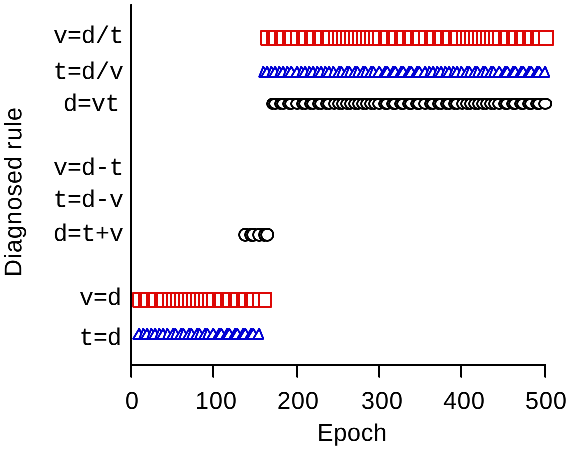

Figure 3. Diagnosis of velocity, time, and distance rules in a static network with two hidden units.

Acquisition of Personal Pronouns

Cascade-correlation networks have also been successfully applied to the acquisition of the English personal pronouns, *me* and *you*. Many children acquire these pronouns without notable errors, whereas other children show persistent reversal errors in which they refer to themselves as *you* and to others as *me*. Psycholinguistic research has shown that the presence of such reversal errors is related to the lack of opportunity to overhear speech that is not addressed directly to the child (Oshima-Takane 1988). Some of this research involved experiments with the so-called *me/you* game (Oshima-Takane 1988). During such a game, a mother might address her child, point to herself, and say *me*. This is an example of speech addressed directly to the child, what we call *addressee* speech. Alternatively, the mother might address the father, point to the father, and say *you*. This is an example of speech overheard by the child, but not addressed to the child, what we call *non-addressee* speech. Some families were assigned to play the me/you game with only addressee speech, whereas other families played with mostly non-addressee speech. A critical question involves the child's productive use of these personal pronouns. What does the child say as he or she joins in the game, when pointing to a parent or to the self?

Results indicated that 19-month-olds who were about to learn personal pronouns profited more in pronoun production from overheard, non-addressee speech than from speech directly addressed to them. In fact, only those children who had opportunities to hear pronouns in non-addressee speech could produce pronouns without errors. Listening to non-addressed speech provides experience with shifting speech roles that is essential to learning the semantics of personal pronouns (Oshima-Takane 1988).

This research was extended to a naturalistic study in which second-borns were found to acquire these pronouns earlier than first-borns, even though these children did not differ on other language measures such as mean length of utterance (Oshima-Takane, Goodz, & Derevensky 1996). Presumably, second-born children have relatively more opportunities to hear pronouns used in speech that is not addressed to them, during conversations between a parent and older sibling.

All of these regularities were simulated with cascade-correlation networks (Shultz, Buckingham, & Oshima-Takane 1994; Oshima-Takane, Takane, & Shultz 1996). The networks were trained to predict the correct pronoun as output given input information on the speaker, addressee, and referent. These networks were trained in two phases, mimicking the me/you game. In the first phase, the network was exposed to speech uttered by parents or other adults. This first phase of training was biased in favor of different amounts of addressed or non-addressed speech. Two particularly interesting conditions reconstructed the language environments of first- and second-born children, respectively. We reasoned that first-borns hear a preponderance of addressee speech from a single caretaker and bit of non-addressee speech in the evening when the parents are together. We implemented this with a ratio of addressee to non-addressee speech of 9:1. Second-borns, in contrast, are likely to hear both addressee and non-addressee speech in about equal measures all day long, as they listen to conversations between the caretaker and their older sibling and are often addressed by those two speakers. We implemented this with a ratio of addressee to non-addressee speech of 1:1.

In the second phase, it was the network's turn to speak, taking the role of a child playing the me/you game. The question was how long the network would take to learn correct pronoun use when addressing others, as a function of the amount of addressee or non-addressee speech it had experienced during the first phase of training.

Results for part of one simulation are shown in Figure 4 in terms of mean epochs to learn each of the two phases under first- and second-born environments. Of most interest is the fact that networks learning phase 1 under second-born conditions, in which addressee and non-addressee utterances were present in equal measure, had a much easier time with the child speaking patterns of phase 2 than did the networks that learned phase 1 under first-born conditions, in which there was a preponderance of addressee speech. Indeed, all of the so-called second-born networks showed immediate, error free generalization to the child speaking patterns of phase 2 under the conditions represented here. In contrast, all of the so-called first-born networks required substantial phase 2 training to reach the same level of pronoun competence.

Both groups of networks succeeded in learning the rules underlying correct pronoun use, namely that a first person pronoun refers to the person using it and a second person pronoun refers to the person who is addressed when it is used. Moreover, the networks were sensitive to the type of speech in the training patterns, such that acquisition was error-free in the case of equal amounts of addressed and non-addressed speech, or characterized by persistent reversal errors in the case of a predominance of directly addressed speech. Errorless generalization was particularly evident when networks could overhear speech involving a number of other people, say an aunt and uncle in addition to the parents, as was true of the conditions represented in Figure 4 (Oshima-Takane, Takane, & Shultz 1996).

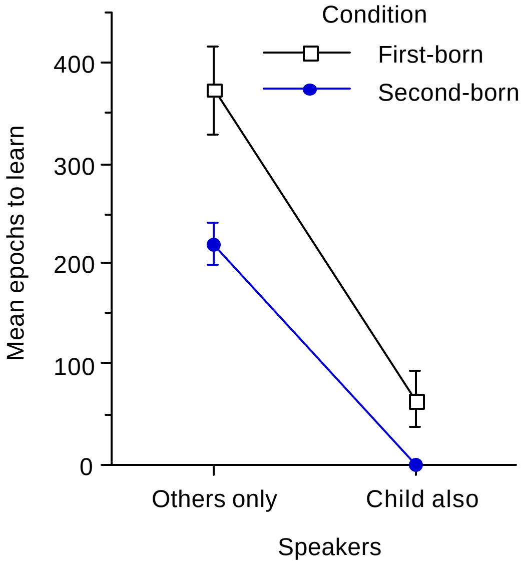

Figure 4. Mean number of epochs to learn pronoun patterns in each of two phases where the first phase (Others only) implements the language environments of either first- or second-born children.

Other Simulated Phenomena and Source of Simulation Power

The foregoing case studies demonstrate two important sources of power in cascade-correlation simulations: gradual increases in computational resources, and sensitivity to environmental bias. In this section, we provide a more general characterization of simulation power in cascade-correlation network models.

Generally speaking, cascade-correlation models succeed in capturing rulelike performance, stage transitions, stage sequences, and a variety of perceptual effects associated with cognitive development in children. Although rules characterizing human performance are not coded explicitly in neural networks, the networks come to behave as if they were following such rules. This rulelike performance derives from the ability of neural networks to extract statistical regularities from the learning environment. In some cases, as for example with personal pronouns, these rules are firm enough to achieve errorless generalization to novel patterns. Furthermore, networks show the capacity for integrating rulelike behavior and performance on exceptions in a smooth homogeneous fashion (Plunkett & Marchman 1991). The rules for integrating velocity, time, and distance cues and those for use of personal pronouns are learned and represented by networks in this implicit way.

Transitions between stages are accomplished in cascade-correlation networks by quantitative connection weight adjustments and qualitative changes in network topology due to hidden unit recruitment. Among the desirable psychological properties of these transitions are their tentative nature (indicated by partially overlapping rule diagnoses in Figure 1) and occasional stage skipping and regression to early stages (not shown here). All of these properties reflect the dynamic, chaotic quality of connectionist networks. Due to the randomness of starting configurations of connection weights and the unpredictability of precise trajectories through weight space and topology space, emerging network behaviors are not rigid.

Stages in children's cognitive development typically occur in particular sequences. The ability of cascade-correlation networks to capture correct stage sequences is due to a variety of factors. In the case of performance on the often simulated balance scale, it is critical for a network to be in a particular region of connection weight space early in its developmental history and to recruit a small number of hidden units (Shultz, Mareschal, & Schmidt 1994; Shultz et al. 1995). For the balance scale task, some number of weights are placed at various distances on the left and right of a fulcrum of a rigid beam, and the child is asked to predict which way the beam will tip when supporting blocks are removed. Children progress through four rulelike stages on the balance scale: (1) use weight information alone, (2) use mainly weight information, but use distance information when the weights on each side are equal, (3) use both weight and distance information equally, but resort to guessing when weight and distance cues give conflicting predictions, and (4) predict correctly no matter what the arrangement of weights (Siegler 1981).

One way for networks to enter that particular critical region of connection weight space early in learning is to learn about balance scales in a environment that is biased in favor of equal distance problems -- problems in which weights are placed at equal distances to the left and right of the fulcrum (McClelland 1995). From such an environment, the network learns that the amount of weight is a much more important predictor of balance scale results than is distance from the fulcrum. This ensures that the network progresses through early stages that emphasize use of weight information.

Network stages in the seriation (ordering) of different sized objects result from a modularization of the seriation task into selecting versus moving an item and slight environmental biases in favor of smaller, less disordered arrays (Mareschal & Shultz 1993). Seriation stages in children involve a progression through four rulelike stages: (1) random moves, (2) partial sorts, (3) complete, but error prone, sorts, and (4) complete systematic sorts with very few errors (Inhelder & Piaget 1969). With the foregoing architectural and environmental constraints, cascade-correlation networks progress through these four stages as well.

Cognitive developmental phenomena are often accompanied by a variety of perceptual effects. An example is the item size effect in seriation, wherein performance improves as size differences between items increases. Such perceptual effects can be expected whenever two or more quantitative values are mapped onto a qualitative comparison. Such perceptual effects are pervasive in cognitive developmental research, but no past theoretical account integrates them with the cognitive features of the task. In neural networks, these perceptual effects are a natural result of the continuous nature of network computations. Larger differences in inputs produce clearer activation patterns on hidden units and more decisive qualitative decisions on output units. When quantitative inputs are reduced to qualitative decisions, some information is inevitably lost. The larger the relevant quantitative differences are, the more accurate the qualitative judgments will be.

Conclusions

As with children, development in cascade-correlation networks can be attributed to a combination of intrinsic and extrinsic factors. In particular, as illustrated by the two cases featured here, it is critical that networks grow in computational power and that they are sensitive to environmental biases. The importance of network growth was established by simulation experiments in which generative networks were compared to static networks. Generative cascade-correlation networks captured the correct sequence of psychological stages, but static networks did not. The importance of environmental bias was established by simulation experiments in which the relative frequencies of various types of training patterns were varied in correspondence with variation in children's naturalistic environments. Cascade-correlation networks responded to these environmental variations in the same way that children do. It is also important that networks are able to abstract regularities from the environment to achieve rulelike behavior and compute unit activations in a continuous manner to simulate perceptual effects.

Acknowledgments

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Buckingham, D., & Shultz, T. R. 1994. A connectionist model of the development of velocity, time, and distance concepts. Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, 72-77. Hillsdale, NJ: Erlbaum.

Buckingham, D., & Shultz, T. R. 1995. Computational power and realistic cognitive development. Technical Report No. 1995, McGill Papers in Cognitive Science, McGill University, Montr al.

Fahlman, S. E. 1988. Faster-learning variations on back-propagation: An empirical study. In D. S. Touretzky, G. E. Hinton, and T. J. Sejnowski (Eds.), Proceedings of the 1988 Connectionist Models Summer School, 38-51. Los Altos, CA: Morgan Kaufmann.

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

Inhelder, B., & Piaget, J. 1969. *The early growth of logic in the child*. New York: Norton.

Mareschal, D., & Shultz, T R. 1993. A connectionist model of the development of seriation. Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society, 676-681. Hillsdale, NJ: Erlbaum.

McClelland, J. L. 1995. A connectionist perspective on learning and development. In T. J. Simon & G. S. Halford eds., *Developing cognitive competence: New approaches to process modeling,* 157-204. Hillsdale, NJ: Erlbaum.

Oshima-Takane, Y. 1988. Children learn from speech not addressed to them: The case of personal pronouns. *Journal of Child Language* 15: 95-108.

Oshima-Takane, Y., Goodz, E., & Derevensky, J. L. 1996. Birth order effects on early language development: Do second born children learn from overheard speech? *Child Development* 67.

Oshima-Takane, Y., Takane, Y, & Shultz, T. R. 1996. The learning of first and second person pronouns in English: Network models and analysis. Submitted for publication.

Plunkett, K, & Marchman, V. 1991. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition* 38: 43-102.

Shultz, T. R., Buckingham, D., & Oshima-Takane, Y. 1994. A connectionist model of the learning of personal pronouns in English. In S. J. Hanson, T. Petsche, M. Kearns, & R. L. Rivest eds., *Computational learning theory and natural learning systems, Vol. 2: Intersection between theory and experiment*, 347-362. Cambridge, MA: MIT Press.

Shultz, T R., Mareschal, D., & Schmidt, W. C. 1994. Modeling cognitive development on balance scale phenomena. *Machine Learning* 16: 57-86.

Shultz, T. R., Schmidt, W. C., Buckingham, D., & Mareschal, D. 1995. Modeling cognitive development with a generative connectionist algorithm. In T. J. Simon & G. S. Halford eds., *Developing cognitive competence: New approaches to process modeling,* 205-261. Hillsdale, NJ: Erlbaum.

Siegler, R. S. 1981. Developmental sequences between and within concepts. *Monographs of the Society for Research in Child Development* 46 (Whole No. 189).

Wilkening, F. 1981. Integrating velocity, time, and distance information: A developmental study. *Cognitive Psychology* 13: 231-247.

Wilkening, F. 1982. Children's knowledge about time, distance, and velocity interrelations. In W. J. Friedman ed., *The developmental psychology of time*, 87-112. NY: Academic Press.