Interactions between Frequency Effects and Age of Acquisition Effects in a Connectionist Network

Paul W. Munro (munro@sis.pitt.edu)
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260 USA

Garrison Cottrell (gary@cs.ucsd.edu)
Department of Computer Science and Engineering 0114
University of California, San Diego
La Jolla, CA 92093-0114 USA

Abstract

The performance of a connectionist network, in which some resources are absent or damaged is examined as a function of various learning parameters. A learning environment is created by generating a set of random "prototypes" and clusters of exemplar vectors surrounding each prototype. An autoencoder is trained on the patterns. The robustness of each learned item is measured as a function of the time at which it was "acquired" by the network and its overall frequency in the environment. Both factors are shown to influence robustness under several learning conditions.

Introduction

For all their shortcomings, feed-forward network models of learning and memory share certain important features with their biological counterparts. Among these are the ability to gradually abstract statistical regularities from their environments by incorporating them into their connectivity structures and the feature generally known as "graceful degradation".

In this paper, the relationship between early learning (acquisition) and degradation of performance through loss of resources is examined in the context of small-scale simulations, in terms of frequency effects, age of acquisition (AoA) effects, prototype effects, and the insertion of noise into the neural network.

The relative influence of AoA compared to frequency on word naming tasks has been argued among cognitive psychologists and linguists for several years now (Brown & Watson, 1987; Morrison et al., 1992; Gerhard & Barry, 1998). Of course, teasing apart the influences of AoA and frequency is confounded by the strong correlation between them. AoA effects have also been reported in other domains, such as object identification and face recognition (Moore & Valentine, 1999). The effects of AoA and frequency on pattern error have been analyzed by Smith, Cottrell, and Anderson (2001). Here, we look at pattern performance in the face of damage to the network, simulating neuronal failure as could occur with aging or trauma.

The robustness of network performance to hidden unit damage has been shown to improve for networks trained with noise among the hidden units (Judd & Munro, 1993). In some cases, this kind of noise has been shown to improve the generalization properties of a network (Clay & Sequin, 1990). Functionally, the hidden representations of the training items settle to states that are further apart in terms of a Euclidean measure.

In this paper, we examine the following three hypotheses:

1. The robustness of an item under loss of network computational resources (analogous to the loss of neurons in humans) is related both to the time at which that item was "acquired", and to the average frequency of the item in the network's experience.

2. Prototypical items are more robust than exemplars, even if they are never explicitly presented to the network, since they share features with populations of exemplars, and thus have high "effective frequencies" in the environment.

3. Early explicit learning of prototypes can result in a more robust set of internal exemplar representations.

Methodology

The training set

A two-step process is used to generate a structured set of bit strings of length L. First, a set of N prototype strings is produced by generating 0 and 1 values

independently with probability 0.5 for each bit having a value 1. In the second step, a set of n_i exemplar strings are generated from the i^{th} prototype P_i by "flipping" bits with a low probability. The result is a set of N pattern "clusters" (see Figure 1). While the network is trained on the exemplar patterns only, the network performance is measured for both the exemplars and the prototypes. In this study, L=100, N=10, and n_i=10, (i=1..10).

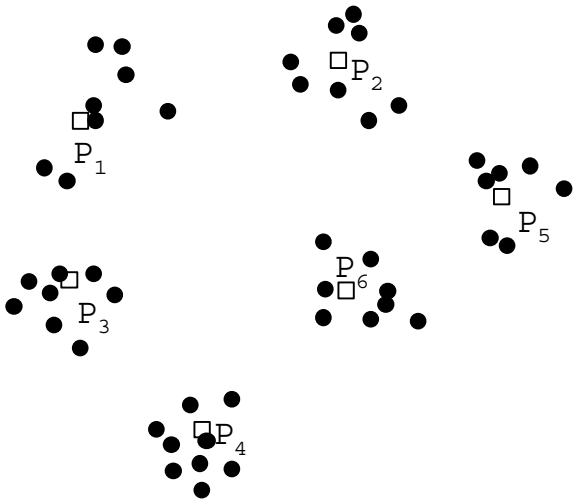

Figure 1. Schematic view of the training patterns. Prototypes (open squares) are randomly generated binary strings. A set of exemplars (filled circles) is generated in the neighborhood of prototype.

In order to analyze the role of frequency, items were selected from the training set according to a ramp distribution; that is, the selection probabilities for the 100 exemplars ranged from approximately 0.0002 to 0.0200 linearly. The probabilities were assigned such that the cluster probabilities also followed a ramp distribution. In other words, the items were ordered according to the parent prototypes and the probability of selecting the k^{th} item was proportional to k. This way, the clusters were also ordered such that the probability of selecting an item from the j^{th} cluster was proportional to j.

Network Architecture

Networks trained by backpropagation to reconstruct their input pattern at the output layer (autoencoders) with a single hidden layer of 40 units are trained using backprop. In some trials, the responses of the hidden units are randomly perturbed to analyze the effect of network noise. An output unit's response is deemed "correct" if it differs from the target by less than a predetermined tolerance level δ. Performance is measured in terms of the number of correct output units. If the network responds with a sufficient number of correct output units to an input pattern, that pattern has been acquired by the network. The point in training at which a pattern is first acquired is called its age of acquisition (AoA). Preliminary studies have shown that in some cases a pattern may briefly be "forgotten" soon after its initial acquisition. In such instances, the forgotten pattern is promptly reacquired; thus, the AoA is defined as the time the pattern is first acquired.

Performance Analysis

After training, the network's response to each training pattern was tested under various damage levels. Damage was implemented by only allowing the output of k of the H hidden units to stimulate the output layer, where k is varied from 1 to H. The minimum number of hidden units required to reconstruct the input pattern (to within a specified degree of tolerance) is recorded as a measure of the pattern's robustness in the network. In some cases, patterns were "forgotten" after initial acquisition. In most such cases, the pattern was reacquired, but not always.

Experimental Conditions

In all the experiments, the acquisition criterion is that 95 out of 100 units should be within 0.2 of their target value (0 or 1). The total training time is either 50000 or 100000 pattern presentations, depending on the condition. Thus, with the ramp distribution, the number of presentations of each individual pattern varies from about 10 to about 2000.

Control Condition (CC) In the control condition, the network is trained with just the 100 exemplars for a period of 100000 pattern presentations.

Head Start Condition (HC) Here, the training set consists a subset of only 10 patterns (one from each cluster) of the full set of 100 exemplars for the first 10000 time steps. This is done to guarantee very low AoAs on some patterns. The training set is expanded to the full set, including the initial subset, for 90000 more presentations. Ellis & Lambon-Ralph (2000) found strong AoA effects in a staged learning condition of this kind.

Noisy Condition (NC) This condition is the same as the previous condition (HC) with "Boolean" noise injected into the hidden layer during the early phase. Here, the activity levels of a small number of hidden units are multiplied by −1. This manipulation is predicted to increase the overall robustness of the full training set.

Prototype Condition (PC) In this variation of HC, the network is trained on only the protoypes during the early phase with no injected noise. Note that prototypes

are never explicitly presented in the previous three conditions.

Results

All conditions show a strong dependence of AoA on frequency. In general, prototype patterns are acquired earlier than exemplar patterns, even if they are not explicitly presented, with the AoA of the prototypes dependent on average frequency of the corresponding exemplars.

Control Condition (CC)

Over the course of 100000 exemplar pattern presentations, 92 of the 100 exemplars were acquired by the network. The eight nonacquired exemplars were all among the 11 least frequent. Of the 10 prototypes, one was not acquired, and eight were acquired in the first 10000 iterations. A scatterplot of AoA vs frequency follows a hyperbolic trend (Fig 2, top). This observation prompted a second scatterplot (Fig 2, bottom), in which AoA is examined vs. $freq^{-1}$. Regression on these data indicates the product of AoA and frequency is about 190 (zero intercept assumed).

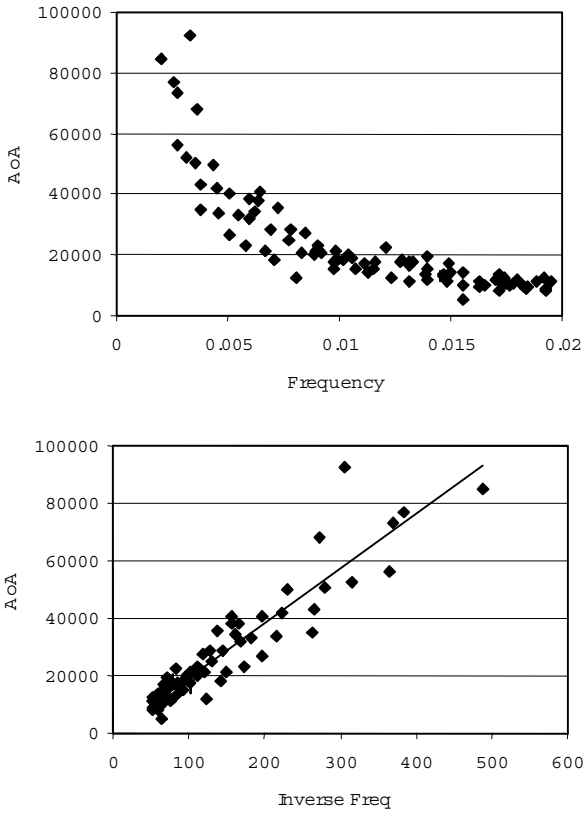

Figure 2: AoA vs Frequency (top) and AoA vs. $Freq^{-1}$ (bottom). The random selection of stimuli in the simulation follows a ramp distribution to give a wide range of frequencies.

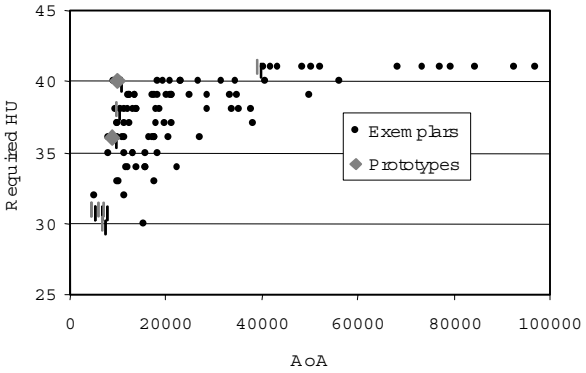

Figure 3. The number of hidden units required to reconstruct the input as a function the AoA. A value of 41 indicates that when the simulation halted, the pattern could not be reconstructed with all 40 hidden units.

The fragility of each item, as measured by the number of hidden units required to reconstruct the pattern tends to be higher for the patterns with later AoA (i.e., earlier patterns are more robust). This is true for both the exemplars and the prototypes (Fig 3). Similarly, items that are more frequent tend to be more robust (Fig 4).

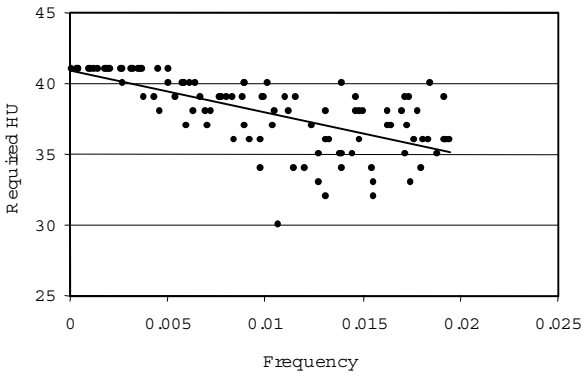

Figure 4. The required number of hidden units vs. frequency. The trendline shows that more frequent items tend to be more robust.

Regression against both variables indicates that the influence of AoA ($p = 0.01139$) is stronger than frequency ($p = 0.03271$) by a factor of almost three.

Head Start Condition (HC)

Here, the first 5000 iterations use only a subset of 10 items (one exemplar from each prototype's "cluster") for training. The network is then exposed to the entire set of 100 exemplars for 45000 subsequent learning trials. Selection of patterns during early exposure also follows a ramp distribution, giving a variety of frequencies within this set.

Early Items. Nine of the 10 items presented alone for the first 5000 time steps are learned before presentation 2000. Four of them are acquired before the earliest prototype (1000 iterations). The least frequent item in this set was never learned. As in CC, AoA and frequency are highly correlated.

Prototypes. The mean AoA for prototypes under HC (12907) is later than it is under CC (10568) and the average prototype is slightly less robust under HC (35.75 HU) than under CC (34.22 HU).

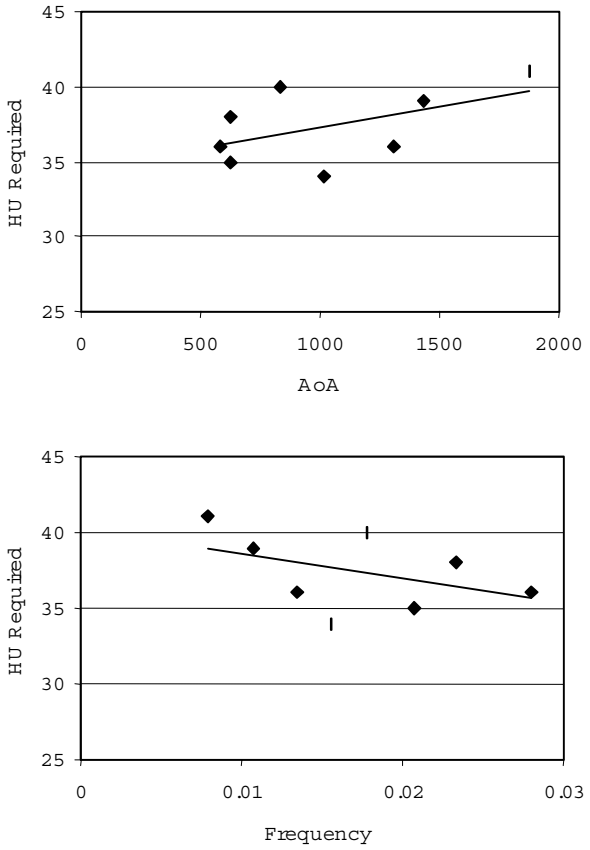

Figure 5. The dependencies of robustness on AoA (top) and frequency (bottom) under HC.

Prototype Condition (PC)

This condition is like HC, except that the ten patterns presented in the early phase are the prototypes of the later patterns. No significant differences in the effects on robustness or AoA were observed in the PC relative to HC.

Noisy Condition (NC)

As in the case of PC, this condition produced mainly negative results. No significant effect of the noise was noticed on the acquisition or robustness of the exemplars. The main observed effect of noise is that the prototypes are acquired much faster. However, the network does not maintain the ability to reconstruct prototypes from the low frequency clusters. Nevertheless, those prototypes that are maintained can withstand more damage to the network.

The bar graphs in Figure 6 display the AoA and robustness (HU required) for the prototype patterns, such that they can be compared with corresponding values in the control condition (black bars=NC, striped bars=CC).

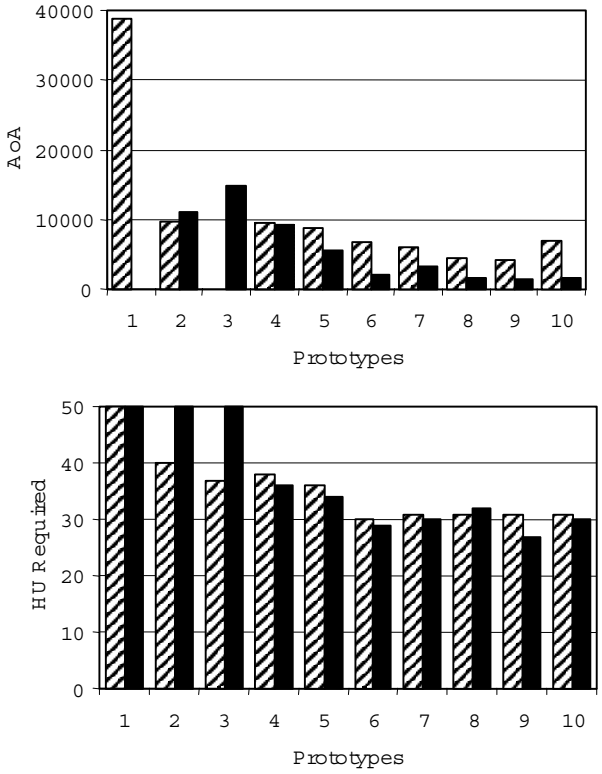

Figure 6. Distributions of AoA (top) and the required number of hidden units (bottom) are displayed for the control condition (striped bars) and NC (dark bars). The 10 items are ranked from lowest (approx. 0.018) to highest frequency (approx. 0.18). The network was never able to reconstruct the lowest frequency prototype (#1), hence there is no bar for this condition. The maximum value for the lower bar graph is the total number of hidden units, 40. A value of 50 means that the network could not reconstruct those prototypes at the end of the simulation.

Conclusions

As a preamble to the data analysis, the relationship between AoA and frequency was examined. These variables were found to be strongly related by a function of the form $a=k/f$, where a is the AoA, f is the frequency, and k is a constant (refer to Fig 2). Even

though this did not bear directly on the hypotheses, it may be the strongest result of this paper!

Our results support the first two hypotheses. The first hypothesis, that both frequency and AoA influence robustness of a learned item is evident from the simulations. Bivariate regression of the robustness variable (HU required) against the two independent variables gave fits that were not very tight (i.e., the p values were too high for the results to be considered significant). Nevertheless, the value corresponding to AoA was consistently lower than that for frequency, indicating a stronger dependence of robustness on AoA.

The second hypothesis, that prototypes are more robust than exemplars was supported by the simulations. The effect is as strong as expected by the measure used here: under CC, prototypes require an average of 34.3 HU, while exemplars require 36.3 HU. Note that this may simply by a byproduct of the AoA effect, since prototypes are acquired much earlier than exemplars. Frequency also plays a role. Even when the prototypes are not explicitly presented, and thus have no frequency per se, the exemplars may be considered distorted versions of the prototypes. Hence, each prototype has an "effective frequency" that depends on the total frequency of its supporting exemplars weighted by the exemplar-prototype distances.

Our simulations did not support the third hypothesis, that early explicit prototype training would result in representations that are more robust. While no such effect has yet been observed, it remains as a subject for future investigation.

Discussion

The issues investigated in this study are the first steps into the exploration of a broader question: How does the adult cognitive structure ultimately depend on the initial stages of learning? This question is quite similar to the age-old debate of nature vs. nurture. Here the issue is whether some potential for later cognitive capabilities is dependent, not on innate factors, but on the content of early experience and the biological mechanisms at work.

The process of acquisition of information, the sequence in which items are presented to the learner, as well as the internal parameters of the learner, may play a determining role in the adult conceptual architecture. It may be that the representations of concepts acquired in childhood, and the associations formed among them construct a foundation on which later concepts are built. Hence, the soundness of this foundation may determine the ultimate robustness of the adult.

Certainly, the importance of early learning on cognitive development has been acknowledged (for example, Catherwood, 1999). In the present work, we have begun to examine this within the connectionist framework, whereby adult cognitive performance might be linked to the statistics of the learning environment in early childhood.

.

Acknowledgments

We would like to thank the members of the GURU group at UCSD for valuable discussions. Paul Munro gratefully acknowledges the hospitality of the group and its leader, coauthor Garrison Cottrell, during a sabbatical leave spent at UCSD during the fall quarter of 2000.

References

Brown, G. & Watson, F. (1987) First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and naming latency. Memory & Cognition, 15:208-216

Catherwood, D. (1999) New Views on the Young Brain: offerings from developmental psychology to early childhood education. Contemporary Issues in Early Childhood, 1:23-35.

Clay, R. & Séquin, C. (1992) Fault tolerance training improves generalization and robustness. In: Proceedings of the International Joint Conference on Neural Networks. IEEE/INNS: Baltimore MD.

Ellis, A. W., & Lambon-Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. Journal of Experimental Psychology: Learning, Memory, & Cognition, 26(5), 1103-1123.

Ghiselli-Crippa, T. & Munro, P. (1994) Emergence of global structure from local associations. In: Advances in Neural Information Processing Systems 6, J.D. Cowan, G. Tesauro, J. Alspector, eds. San Mateo, CA: Morgan Kaufmann.

Judd, S. & Munro, P. (1993) Nets with unreliable hidden nodes learn error-correcting codes. In: Giles, C.L., Hanson, S.J., Cowan, J.D., (eds.) Advances in Neural Information Processing Systems 5. San Mateo, CA: Morgan Kaufmann Publishers.

Moore, V. & Valentine, T. (1999) The effects of age-of-acquisition in processing famous faces and names: Exploring the locus and proposing a mechanism. Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society. Mahwah NJ: Erlbaum.

Smith, M., Cottrell, G., and Anderson K. (2001) The early word catches the weights. To appear in: Advances in Neural Information Processing Systems 12 MIT Press, Cambridge, MA.