Proceedings of the Annual Meeting of the Cognitive Science Society

Neural Network Models of Memory Span

Proceedings of the Annual Meeting of the Cognitive Science Society, 11(0)

Schweickert, Richard
Guentert, Lawrence
Hersberger, Lora

1989

Neural Network Models of Memory Span

Richard Schweickert

Lawrence Guentert

Lora Hersberger

Department of Psychological Sciences

Purdue University

Abstract

A model is presented in which short term memory is maintained by movement of vectors from one layer to another. This architecture is ideal for representing item order. Two mechanisms for accounting for serial position curves are considered, lateral inhibition, and noise from neighboring items. These also account for effects of grouping by inserting pauses during presentation. Two other effects, a reverse word-length effect and the effect of phonological similarity, are attributed to the reconstruction of items from partially decayed traces. If all the phonemes in an item are intact at recall, the item is recalled correctly. Otherwise, the subject guesses according to a model developed by Paul Luce for identification of words presented in noise.

Introduction

Several methods for short term memory storage have been proposed for neural networks, including quickly decaying weights (Hinton & Plaut, 1987), sustained activation, and moving activation (Hebb, 1949). Fast weights are useful for maintaining temporary learning, but not for the roughly 2 second duration of immediate memory (Mackworth, 1963). Sustained activation is possible, but it seems unlikely that it would not spread to neighboring units. Activation moving from layer to layer is not only likely, but accounts for several phenomena of short term memory in a natural way.

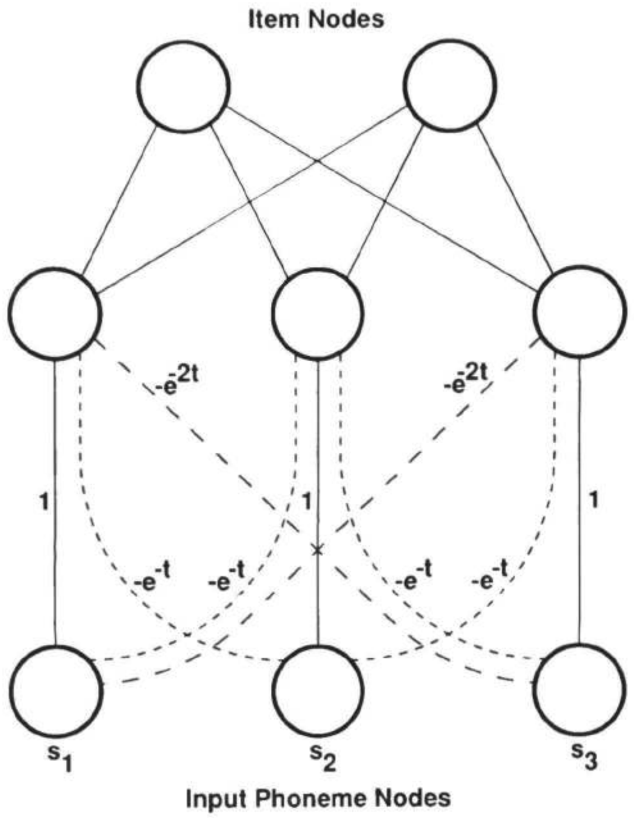
Memory span for a type of item, such as digits, is the number of items that a subject can immediately recall in order half the time. A key finding is that many errors are transpositions, rather than omissions or substitutions. The usual primacy and recency effects found in free recall are also found in immediate recall. Moreover, if there is an empty time interval between two items at presentation, there is a recency effect prior to the gap and a primacy effect after the gap (e. g., Huttenlocker & Burke, 1976). This suggests interference between items presented close together in time.

Interference could be explained by lateral inhibition of temporally adjacent items. Primacy and recency would occur because items at the extremes receive inhibition only from neighbors on one side. Recently, a way of accounting for effects such as Mach bands in vision without lateral inhibition has been proposed (Cornsweet & Yellot, 1985). With constant volume operators, excitation spreads from each unit, with the extent of the spreading inversely proportional to the intensity exciting the unit. Edge enhancement is one result. Yet another proposal for interference is that the positions of items are perturbed, so items are recalled in the wrong order (Lee & Estes, 1981). Items at the extremes can only move in one direction, and so are less likely to move.

If excitation spreads to neighboring items, excitation favoring an item could be strongest at a position different from the one the item was presented in. Hence, perturbations can be explained in terms of spreading excitation. A further simplification of the possibilities is that when lateral inhibition is combined with constant volume operators, erroneous predictions are made for vision (Yellot, 1989). The explanations are likely to be mutually exclusive for memory also. It is difficult to distinguish the effects of lateral inhibition from those of constant volume operators, even in vision (Yellot, 1989, p. 33), so we will present an example of a model of each kind.

A Lateral Inhibition Model

A model based on lateral inhibition is shown in Figure 1. It explains primacy and recency effects, and the effects of temporal gaps in presentation. When a phoneme i is presented to the input layer, its strength is s_i. After time t its strength decays to $s_i \exp(-t)$. The node in the output layer corresponding to phoneme i receives excitation from the input node for phoneme i, and receives inhibition from all the other nodes in the input layer. The closer the time of presentation of phoneme j is to the presentation of phoneme i, the greater the inhibition between them. Let t_{ij} be the time elapsing from presentation of i to presentation of j (regardless of which came first). For simplicity, assume the onset to onset interval for each phoneme is the same, t. If k phonemes intervene between phonemes i and j, then $t_{ij} = kt$. The inhibition of item j on item i is $s_j \exp(-t_{ij})$. The output of the node for phoneme i in the memory layer at time t is the log odds of correctly identifying phoneme i, based on the trace itself.



<u>Figure 1</u>. Phonemes in the bottom layer send excitation and inhibition to nodes in the next layer. Inhibition between two phonemes decays as a function of the time intervening between them.

Strengths in the memory layer also decay exponentially with time. These are the input to the word layer at the time or recall. Then at a time T,

$$\log [p_i(T)/(1 - p_i(T))] = (s_i \qquad s_j \exp(-t_{ij}))) \exp(-T).$$

Here $p_i(T)$ is the probability that phoneme i is recalled correctly at time T.

For simplicity, suppose the time to read each word is the same as the time to recite it during recall. Then it is easy to see that the time elapsing between presentation of phoneme i and its recall is L, where L is the total time to pronounce all the items in the list.

A Spreading Excitation Model

Consider an array of rows and columns of neural units. A phoneme is a column vector, where each component is the strength of some feature. When a phoneme is input to the first column of the array, the excitation in a row is transferred to the next column, and to the next and so on. Suppose the time required for transfer is shorter, and has smaller variance, the stronger the excitation. When excitation spreads from a column to an adjacent column, the original excitation remains, but decays over time. After a while, the excitation for a given feature will be spread over a set of columns. If excitation from features of two or more phonemes arrives at the same column, the excitations are summed. This mode of spreading is close enough to that of constant volume operators to produce the "edge enhancement" analogous to primacy and recency. Items at the extremes only have neighbors on one side, so they have less noise added to their representations.

It is plausible that excitation would spread, and at different rates for different features. This gives a mechanism for assumptions of some models in the literature. First, the perturbations of Lee & Estes (1981) occur because excitation for one item may spread faster than that for another, so a later item may overtake an earlier one. Second, according to Glenberg and Swanson (1986), visually presented items are less distinct in terms of temporal order than auditorially presented items, resulting in better recall for the auditory items (the modality effect). The difference in temporal distinctiveness may be due to differences in the extent of the spread of excitement. Finally, the TRACE model for speech recognition by McClelland and Elman (1986) proposes that the activation for entities is spread over neighboring units. Their model is static, but a snapshot of the model sketched above would look like the neural array in the TRACE model.

Primacy, recency, and grouping effects are due to the mechanisms by which traces deteriorate. A subject processing a partially degraded trace will try to reconstruct the original item. The next two effects to be discussed are explained in terms of reconstruction.

The Reverse Word-Length Effect

Ordinarily, memory span is shorter for items taking longer to pronounce (the word-length effect). As a rule, the memory span for a type of item is the length of a list of such items that can be pronounced in about 2 seconds (Baddeley, Thomson, & Buchanan, 1975; Mackworth, 1963; Schweickert & Boruff, 1986). The span is slightly greater for familiar items. To learn about the role of familiarity, we investigaged memory span in highly practiced subjects.

Two subjects completed 30 sessions in a memory span experiment. There were twenty items in each of five types of item. The lengths of the lists were from 3 to 9 items. At the beginning of each trial, a list appeared on a CRT. Subjects read the list aloud, and speaking durations were measured. Subjects then tried to recall the list in order.

Each subject produced a reverse word length effect, that is, the slower the speaking rate in items per second, the more items that could be recalled in a given time period. The data for subject 1 are in Table 1, the other subject's data are similar.

Table 1.

Material	letters	words	prepositions	colors	shapes
Rate	3.11	2.77	2.75	2.62	1.94
Recall	50%	60%	69%	78%	90%

The Phonological Similarity Effect

In a pronunciation task, Chase (1975) found that subjects pronounce phonologically similar items more slowly than dissimilar ones. The difference in speech rate raises an interesting question. Does the rate difference account for the effect of phonological similarity on span?

We carried out an experiment to investigate this question. Phonologically similar lists were made of items from the set {b, c, d, g, j, k, p, t, v, z}. Dissimilar lists were made from the set {b, d, f, h, k, l, m, q, r, z}. Names of letters in the first set all end in long e or long a, making them more similar than those in the second set.

Eighteen subjects served individually for one hour. The session began with a practice block of digits, followed by two blocks each of similar and dissimilar items. Subjects were randomly assigned to six groups, corresponding to the six possible orderings of two similar and two dissimilar blocks.

The span for the similar items was 5.62, that for the dissimilar items was 7.06, a significant difference in an analysis of variance ($p < .001$). The pronunciation rates were almost identical, 3.01 items per second for the similar items, and 2.92 for the dissimilar. In short, when the task is not only to pronounce the items, but to recall them as well, speaking rates are the same for the similar and dissimilar items. The conjecture that the effect of phonological similarity on memory span is due to a slower pronunciation rate for phonologically similar items is not supported.

Redintegration

If the phonemes in item are not all recalled, a guess is made from the set of possible items. Identifying an item from a noisy memory trace is analogous to identifying an auditorially presented word in noise. A model for the latter task was developed by Paul A. Luce (1986). In this model, the probability of correct recall of all the phonemes in a word is the product of the probabilities of correctly recalling the phonemes individually. Further, the probability of a correct guess depends on the phonological similarity of items to their neighbors, and on word frequency. The model seems well suited to the present situation, and will be used here as the mechanism whereby a guess is made at recall of a noisy memory trace.

According to the Luce model, if all the phonemes of a word w are not recalled, then

$$g_w = \frac{h_{ww}\ b_w}{\sum h_{wv}\ b_v} .$$

Here, g_w is the probability of guessing word w correctly and h_{wv} is a parameter which increases as the phonological similarity between words w and v increases. The bias b_v in favor of responding with word v is influenced by word frequency.

Learning

The subject will improve his ability to guess considerably if he can tune his bias to match the actual presentation rates of the items in the experiment. Bush, Luce & Rose (1963) proposed a learning rule with two desirable properties in the limit as the number of trials approaches infinity. First, the probability of responding w approaches the value given in the formula above according to the ratio rule. Second, the biases become proportional to the presentation rates actually used in the experiment.

For every v and w, let $g_{wv,i}$ be the probability of guessing w on trial i, given that v was presented. Suppose on trial i, v′ was presented. Then on trial i + 1, for every v and w

$$g_{wv,i+1} \quad g_{wv,i} = ch_{vv'}[d_{vv'} \quad g_{wv,i}],$$

where c, a constant, is the learning rate, and $d_{vv'}$ is 1 if v = v′ and 0 otherwise.

Matters would be more complicated if the guessing probability depended on word length. However, the first and last phonemes are the most crucial for identifying a word (Garner, 1962) so the number of phonemes in the middle may not have much influence. Empirically, Luce and Pisoni (1986) found a very low correlation between word length and accuracy of word identification.

Correct recall of item i occurs either from the trace directly, with probability $p_w(L)$, or, by guessing. That is,

$$P(\text{recall } w) = p_w(L) + [1 \quad p_w(L)]g_w.$$

Note that $p_w(L)$ is influenced by word length and g_w by phonological similarity. In this way, the model uses different mechanisms for the effects of similarity and word length.

The subject can control the spoken duration of each item, and L is the sum of the durations of all the items on the list. Laugherty (1969) reports that immediate memory performance is not monotonic with presentation rate. The value of L which optimizes the probability of recall of item i does not depend on phonological similarity, however. Therefore, phonologically similar items will be pronounced at the same rate when immediate recall follows.

References

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). *Journal of Verbal Learning and Verbal Behavior*, *14*, 575.

Bush, R. R., Luce, R. D., & Rose, R. M. (1963). Learning models for psychophysics. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*, Vol. 1. Stanford: Stanford University Press.

Chase, W. G. (1975). In S. Dornic (Ed.), *Attention and Performance VI*, Erlbaum, Hillsdale, NJ.

Cornsweet, T. N., & Yellot, J. I., Jr. (1985). Intensity-dependent spatial summation. *Journal of the Optical Society of Amerika-A*, *2*, 1769-1786.

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. New York: Wiley.

Glenberg, A. M., & Swanson, N. G. (1986). A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 3-15.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

Hinton, G. E., & Plaut, D. C. (1987). Using fast weights to deblur old memories. *Program of the Ninth Annual Conference of the Cognitive Science Society*. Hillsdale, HJ: Erlbaum.

Huttenlocker, J., & Burke, D. (1979). Why does memory span decrease with age? *Cognitive Psychology*, *8*, 1-31.

Laugherty, K. R. (1969). Computer simulation of short term memory: A component decay model. In G. Bower & J. T. Spence (Eds.), *The psychology of learning and motivation*, Vol. 3, 135. Orlando, FL: Academic Press.

Lee, C. L., & Estes, W. K. (1981). Item and order information in short-term memory: Evidence for multilevel perturbation processes. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 149-169.

Luce, P. A. (1986). Research on speech perception. Speech Research Laboratory, *Technical Report No*. *6*, Department of Psychology, Indiana University.

Luce, P. A., & Pisoni, D. B. (1986). Paper presented at Psychonomic Society Meeting, New Orleans.

Mackworth, J. F. (1963). The duration of the visual image. *Canadian Journal of Psychology*, *17*, 62-81.

McClelland, J. L., & Elman, J. L. (1986). Interactive processes in speech perception: The TRACE model. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, pp. 58-121. Cambrige: MIT Press.

Schweickert, R., & Boruff, B. (1986). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 419.

Yellot, J. I., Jr. (1989). Constant volume operators and lateral inhibition. *Journal of Mathematical Psychology*, *33*, 1-35.