Spreading Activation in PDP networks

James Hendler*
Computer Science Dept.
University of Maryland
College Park, Md. 20742
hendler@cs.umd.edu

Abstract

One argument in favor of current PDP models has been that the availability of "hidden units" allows the system to create an internal representation of the input domain, and to use this representation in producing output weights. The "microfeatures" learned by sets of hidden units, it has been argued, provide an alternative to symbols for certain reasoning tasks. In this paper we try to further this argument by demonstrating several results that indicate that such representations are formed. We show that by using a spreading activation model over the weights learned by networks trained via backpropagation, we can model certain cognitive effects. In particular, we show some results in the areas of modeling phoneme confusions and handling word-sense disambiguation, and some preliminary results demonstrating that priming effects can be modeled by this activation spreading approach.

1 Introduction

A primary difference between the neural networks of 20 years ago and the current generation of connectionist models is the addition of mechanisms which permit the system to create an internal representation. These "subsymbolic," semantically unnameable, features which are induced by connectionist learning algorithms have been discussed as being of import both in structured and distributed connectionist networks (cf. Feldman and Ballard, 1982; Rumelhart and McClelland, 1986). The fact that network learning algorithms can create these *microfeatures* is not, however, enough in itself to account for how cognition works. Most of what we call intelligent thought derives from being able to reason about the relations between objects, to hypothesize about events and objects, etc. If PDP models are to be used for cognitive modeling we must complete the story by explaining how networks can reason in the way that humans (or other intelligent beings) do.

To be able to claim that the internal representations learned by connectionist networks can provide a substrate on which such symbolic reasoning can be performed, we must be

*Also affiliated with the UM Institute for Advanced Computer Studies and the Systems Research Center. Partial support for this work was provided by the Office of Naval Research contract N00014-88-K-0560 and NSF grant IRI-8907890. This work was performed, in part, at the International Computer Science Institute in Berkeley, Ca.

able to demonstrate that this representation, *when removed from an input mapping,* can allow output concepts to be related together. Thus, for example, to account for the sorts of tasks which have motivated traditional AI models, such as natural language processing and planning systems, a system must be able to reason about "grandma" without having grandma (or someone who looks like her) physically available to the perceiver[1]. In a multi-layer PDP network, this minimally requires the ability to abstract relationships between the output units without directly activating input units.

In humans, this relatedness of concepts is usually tested via experimental paradigms that are based on a model in which activation spreads through an associative network relating concepts which are either perceptually or semantically linked. An examination of the spread of activation through semantically related concepts has been the focus of much work in categorization and lexical access[2]. Could this model, in which activation spreads between related concepts (represented by output concepts), be realized using the sorts of microfeature-based representations (encoded in the "hidden" units) learned by PDP models?

In this paper, we will demonstrate some evidence that this sort of model can, in fact, be realized in PDP models by using a direct analog of spreading activation. Essentially, activation at the output nodes is spread through the network of weights between hidden and output units in a three-layer network trained by the classical error back-propagation method. Using this technique, described in more detail in the next section, we are able to demonstrate that meaningful relationships between output units can be found. Following this we describe three results of this work: some results in the areas of modeling phoneme confusions (section 3) and handling word-sense disambiguation (section 4), and some preliminary results in demonstrating that priming effects can be demonstrated during this activation spreading (section 5).

It should be noted, however, that the evidence presented in this paper should not be taken as a direct model of human cognitive processing and activation spread. The differences between multiple-layer, back propagation trained, PDP networks and the either the hardware of the human brain or the human cognitive apparatus are many. All we wish to do in this paper is to demonstrate that distributed representations, such as those learned by these networks, could account for activation spreading effects. Thus, this work (and numerous extensions thereto) is *necessary* to demonstrating that a subsymbolic substrate could implement human-like cognition; however it is far from sufficient.

2 Activation Spread

Our system starts by assuming an activation pattern is started at one or more of the **output** nodes of the PDP network. This activation then spreads to the hidden unit and back to the output units. In this way, those units which share the most "microfeatures" in common

[1]Similarly, one can test this effect simply by closing one's eyes and thinking about "grandma" or any other symbolic entity.

[2]A complete list of citations is beyond the length limit on this paper. A long discussion of spreading activation models in AI and in psychology can be found in (Hendler, 1987; Chapter 8).

will gain the most activation[3].

The spread of activation between the output and hidden units in a PDP network is easily modeled. Consider a network, already trained, in which we have two output nodes, j and k. If j is activated with some energy, that energy will pass to each hidden node in proportion to the weight between j and that node. Each of these nodes, in turn, pass activation to the output nodes in proportion to the weights to those output nodes. Thus, k will gain activation from j given by:

$$v_{jk} = \sum_{i=1}^{I} w_{ji}.w_{ik},$$

where v_{jk} can be considered as a weight between output units k and i. In these networks, we treat the weights as symmetric, so that $v_{jk} = v_{kj}$. Where this symmetry holds, $V = [v_{jk}], (j,k = 1,...,N)$ (where N is the number of output units) is recognized to be, by definition, the mathematical covariance matrix of the weights between the hidden layer and the outputs. Thus, a traditional view of activation spreading, applied to these networks, is modeled by the well-behaved mathematical relationship of covariance[4]

While the covariance numbers are directly related to the patterns learned by the PDP network, they are unbounded, making them difficult to work with in modeling. As it is preferable to handle bounded numbers, in the experiments described in this paper, we replace the covariance matrix V by the mathematically "equivalent," although bounded matrix of correlation coefficients $C = [c_{kl}]$ computed as:

$$c_{kl} = \frac{\sigma_{kl}}{\sqrt{\sigma_{kk}.\sigma_{ll}}},$$

$$\sigma_{kl} = \frac{1}{I}\sum_{i=1}^{I}(w_{ik} - \overline{w_k})(w_{il} - \overline{w_l}),$$

$$\overline{w_k} = \frac{1}{I}\sum_{i=1}^{I} w_{ik},$$

and, consequently:

$$\begin{cases} c_{kk} = 1 \\ -1 \leq c_{kl} \leq +1. \end{cases}$$

The values c_{kl} are bounded and reflect the correlation, based on the parameters of the network, between output units k and l, $+1$ standing for the maximum positive correlation (i.e. virtually identical classes) and -1 for the maximum negative correlation (i.e. completely different).

In the remainder of this paper, we will examine what happens when this model of activation is applied to several specific networks. We will describe the training of the PDP models, and describe how the correlation-coefficient modeled, activation-spreading process can be used to show interesting aspects of the representations learned by PDP networks.

[3]This effect was first demonstrated in a hybrid system merging a local connectionist model and a symbolic marker-passer. Details of that work can be found in (Hendler, 1989).

[4]Others have modeled the spread of activation through a semantic memory using more complex functions. The best known of this work is the ACT* model discussed in Anderson (1983), which also presents a review of other systems using similar techniques.

3 Phoneme Confusion

Many phenomena needing to be explained by the cognitive scientist have their roots in the perceptual similarities between various objects as viewed by the human cognitive apparatus. Some examples include:

1. *Categorization*, where humans can classify objects as better and worse examples of some category (for example, "sparrow" is consistently rated as a better "bird" than "turkey").

2. Priming where the priming occurs for perceptually similar objects. An example of this is "rhyming priming," reported in numerous lexical access experiments, in which words which rhyme demonstrate priming effects.

3. Functional identification in which objects which have perceptual features in common are used to perform functions associated with each other (for example, substituting a rock for a hammer due to a similarity in mass).

Traditional symbolic modeling, in the AI and cognitive psychology communities, has been unable to offer an explanation of these effects.

In the human perceptual system, one of the "earliest" places in which similarity effects can be seen is in the perception of phonemes in continuous speech. Experimentation (Aubert, 1988) has shown that word confusions may arise from phoneme confusions occuring during speech perception. Thus, for example, the phoneme for the "short a" (Ahh) sound will be more likely to be confused with the perceptually related "short e" (ehh) sound than with the less related "hard g" (Guh) sound. Based on experimentation, Aubert produced a matrix of confusion likelihoods (Bounded from 0 to 1) between each of 50 phonemes he tested (thus producing a 50x50 correlation matrix between phonemes, based on human data). In his matrix, 194 cells had values different than 0 (no confusion) or 1 (the identity cells along the main diagonal of the matrix).

To see whether the activation spreading model described in section 2 would produce a similar matrix of phoneme confusions, an experiment was run in which a PDP network (with input, hidden, and output units) was trained to do phoneme identification. Using a technique developed by Bourlard, Morgan, and Wellekens (1989), a data base consisting of 100 sentences were used for training the network to recognize phonemes. Vector-quantized (132 prototypes) mel cepstra were used as acoustic vectors. To simplify the representation of the input data, each vector was replaced by its index coded by a simple binary vector with only one bit "on". Multiple frames were used as input to provide context (9 frames) to the network. Thus, the input field contained $9 \times 132 = 1188$ units, and the total of possible inputs was equal to 132^9. The size of the output layer was kept fixed at 50 units, corresponding to the 50 phonemes to be recognized. There were 26,767 training patterns representing only a small fraction of the possible inputs. A network with five hidden units was trained on this set, thus the network had 1188 inputs, 5 hidden, and 50 output units.

We compared the matrix of correlation coefficients (i.e. the confusion matrix C) to Aubert's phoneme confusion matrix. For the 194 cases where the hand generated matrix contained non-zero values, the correlation was $\rho = 0.365$. Using the entire confusion matrix except for the identity cases, the correlation coefficient was $\rho = 0.285$. In both cases, there

is a statistically significant (P >0.001) correlation between the hand-generated confusion matrix and the one obtained from the PDP network by the spreading activation model.

4 Word Sense Disambiguation

At a "later" level of cognition, word sense disambiguation models have been proposed to account for lexical access data found in psychological experiments (cf. Swinney, 1979). One such model is a structured connectionist model developed by Gary Cottrell at the University of Rochester (Cottrell, 1985). Cottrell, using weights derived by hand, demonstrated that a structured connectionist network could distinguish both word-sense and case-slot assignments for ambiguous lexical items, in a manner consistent with experimental results. Presented with the sentence "John threw the fight" the system would immediately activate both meanings of "throw," but in a short time would settle in an activation pattern in which a node corresponding with only one of the meanings would remain on. Presented with "John threw the ball" it settle on another meaning. The nodes of Cottrell's network included words (John, Threw, etc.), word senses (John1, Propel, etc.) and case-slots (TAGT (agent of the throw), PAGT (agent of the Propel), etc.).

To duplicate Gary's network via training, we used back propagation to train a network, using a training set in which distributed patterns, very loosely corresponding to a "dictionary" of word encodings[5] were associated with a vector representing each of the individual nodes which would be represented in Cottrell's system, but with no structure. Thus, a typical element in the training set could be, for example, a 16 bit vector (representing a four word sentence, each word as a 4 bit pattern), associated with another 16 bit vector representing the nodes:
Bob1 John1 Propel Threw Fight1 Ball1 Pagt Pobj Tagt Tobj Bob John Threw The Fight Ball

For this example, the system was trained on the encodings of the four sentences:

1. John threw the ball.
2. John threw the fight.
3. Bob threw the ball.
4. Bob threw the fight.

with the output set high for those objects in the second vector which were appropriately associated.

Upon completion of the learning, the activation spreading algorithm was used to derive a table of connectivity weights between the output units. These weights were then transferred into the Rochester Connectionist Simulator (Goddard, et.al., 1987), the same simulation method used by Cottrell, and the activation spreading model was used to examine the results. Using the activation spreading method described in section 2, results similar in time-course and behavior to those produced by Cottrell's model were seen. Thus:

[5]In a realistic , these would be replaced by actual signal processing outputs or other representations of actual word pronunciation forms. This technique of using a random encoding is based on the work of Jeff Elman (1988).

1. Activation from the nodes corresponding to *john, throw, the,* and *fight* cause a positive activation at the node for "Throw" and a negative activation at the node for "Propel."

2. Activation from *john throw the ball* spread positively to "Propel" and not to "throw."

3. Activation at *TAGT* and *TOBJ* spreads positive activation to *Throw* and not to *Propel.*

4. Activation at *PAGT* and *POBJ* causes a spread to *Propel* but not to *Throw.*

(We have also used this approach to test more complex sentences, still within the framework of Cottrell's system. Similar results have consistently been obtained.)

5 Priming Effects

A consistent effect observed in experimentation with humans has been the priming effects that are largely responsible for the belief in an autonomic activation-spreading system[6]. Such effects, however, are not exhibited in even the recurrent PDP models. One particular aspect of these effects is the ability for "semantic" expectations to prime recognition and categorization tasks. For example, when expecting a "vowel," *e* will be more quickly recognized as a letter than if primed to expect a "number." Thus, an activation spreading method should allow prior activation of a concept to facilitate recognition of an example of the concept (vowel/*e*, etc.). This facilitation can appear both in a shortened time course to recognition, or in a preference for a recognition of an ambiguous signal based on an expectation.

We have recently begun experimention which shows that priming effects may be induced via the activation spreading method described in section 2. That is, given a previous activation at a particular node, we may cause some other node to "win" more activation energy, faster, than it would if the previous activation was either missing or was on some other node.

To demonstrate this effect, we trained a 12-4-12 auto-associative network[7] to recognize a training set in which the twelve inputs corresponded to the numbers 0 through 9, and two extra inputs, one of which was on when an odd number was presented, the other on when an even number was presented (we'll call these nodes even and odd for simplicity). Thus, the numeral 3 would be represented as 0's in positions corresponding to other numbers and to even, and 1's in the positions corresponding to the number itself and to odd, that is "0 0 0 1 0 0 0 0 0 0 1 0." After training, on the encodings of all 10 numerals, the covariance coefficients were computed and transferred into the Rochester Connectionist Simulator, as in the previous section.

To test for the ability to simulate priming, ambiguous activation patterns were used to observe network behavior. Thus, the system might have the output corresponding to the numeral 3 activated at a strength of .4, and the numeral 6 represented at a value of .6. As the system settled, one or the other of these nodes would become positive, while the other

[6] A good discussion of these effects and related experiments can be found in Anderson and Bower (1983).

[7] that is, one in which the inputs and outputs in the training set were identical

would become negative. Where no other activation was introduced, the node with higher activation would win out over the other[8].

Priming effects are introduced in these networks by first activating either the node *odd* or the node *even* for a short time, followed by the presentation of the ambiguous input. In this situation, the following sorts of behaviors are seen:

1. Where the direction (odd or even) and the number with the larger activation are the same, that number gains ascendancy (becomes more positive while the other node becomes negative) more quickly than where no prior activation is used.

2. Where the direction and the item with less activation (.4) are different, the item with less activation will end up in ascendancy (as opposed to becoming negative as happens without the presence of the priming activation).

Thus, this activation does correspond well with priming effects. It should be noted that as this technique has only been used on quite small data sets, there is a question as to whether the results will scale for more significant trials. Experimentation in this direction is currently underway.

6 Conclusions

In this paper, we have presented some evidence that the sorts of representations learned, via training, by connectionist networks, may have the necessary properties to be able to demonstrate several effects known to occur in human cognition. Using a spreading activation model over the PDP network, we have shown evidence for the ability to model a simple recognition of perceptually related items (the phoneme confusions) and linking of semantically related items (the lexical items in Cottrell's model). In addition, we've discussed some preliminary evidence that priming effects, a robust phenomena in the activation spreading literature, can be shown in this spreading-activation model. Thus, we have presented evidence demonstrated that distributed representations, such as those learned by these networks, could possibly account for activation spreading effects as is required to account for many known psychological results.

As noted in the introduction, however, the evidence presented in this paper should not be taken as a direct model of human cognitive processing and activation spread. The differences between multiple-layer, back propagation trained, PDP networks and the either the hardware of the human brain or the human cognitive apparatus are many. Thus, the work presented in this paper, and numerous extensions thereto, are *necessary* to demonstrating that a subsymbolic substrate could implement human-like cognition; however it is far from sufficient. We are currently examining the use of this technique in more complex networks that simple feed-forward, completely connected networks, and believe that similar effects must be shown if these networks are to be taken seriously as models underlying human cognition.

[8]We should note, however, that these networks often fail to stabilize, and thus all weights go to 0 after a relatively short time. This is the main reason why we categorize these results as "preliminary."

References

[1] Anderson, J.R. *The Architecture of Cognition,* Harvard University Press, Massachusetts, 1983.

[2] Anderson, J.R. and Bower, G.H. *Human Associative Memory,* Lawrence Erlbaum Associates, New Jersey, 1979.

[3] Aubert, X. (1988). Personal communication.

[4] Bourlard, H., Morgan, N., Wellekens, C.J. (1989c). Statistical Inference in Multilayer Perceptrons and Hidden Markov Models with Applications in Continuous Speech Recognition, to appear in *Neuro Computing, Algorithms, Architectures and Applications*, NATO ASI Series.

[5] Cottrell, G.W. *A Connectionist Approach to Word Sense Disambiguation* Doctoral Dissertation, Computer Science Department, University of Rochester, May, 1985.

[6] Elman, J.L. (1988). Finding Structure in Time, *CRL Tech, Report 8801*, University of California, San Diego,

[7] Feldman, J.A. and Ballard, D.H. Connectionist models and their properties *Cognitive Science,* 6, 1982. *205-254.*

[8] Goddard, N., Lynne, K. and Mintz, T. *The Rochester Connectionist Simulator*, TR233, Dept. of Computer Science, University of Rochester, 1987.

[9] Hendler, J. Marker-passing over microfeatures: Towards a hybrid symbolic/connectionist model *Cognitive Science*, 13(*1*), March, 1989 *p. 79-106.*

[10] Hendler, J.A. *Integrating Marker-passing and Problem Solving: A spreading activation approach to improved choice in planning* Lawrence Erlbaum Associates, N.J., Dec. 1987.

[11] Rumelhart, D.E., McClelland, J.L. and the PDP Research Group *Parallel Distributed Computing, (Volume 1 and Volume 2)* MIT Press, Cambridge, Ma., 1986.

[12] Swinney, D.A. Lexical access during sentence comprehension: (Re)Consideration of context effects *Journal of Verbal Learning and Verbal Behavior,* 18, 1979, *645-659.*