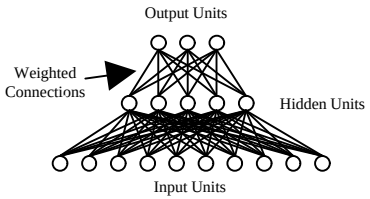
Coarse Coding In Value Unit Networks: Subsymbolic Implications Of Nonmonotonic PDP Networks

C. Darren Piercey & Michael R. W. Dawson
University of Alberta
Edmonton Alberta Canada

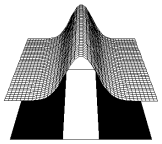
Abstract

PDP networks that use nonmonotonic activation functions often produce hidden regularities that permit the internal structure of these networks to be interpreted (Berkeley et al, 1995; Dawson, 1998; McCaughan, 1997). In some cases, these regularities are associated with local interpretations (Dawson, Medler, & Berkeley, 1997). Berkeley has used this observation to suggest that there are fewer differences between symbols and subsymbols than one might expect (Berkeley, 1997). We suggest below that this kind of conclusion is premature, because it ignores the fact that regardless of their content, the local features of these networks are not combined symbolically. We illustrate this point with the interpretation of a network trained on a variant of Hinton's (1986) kinship problem, and show how the network's behavior depends on the coarse coding of the information represented by hidden unit bands, even when these bands have local interpretations. We conclude that nonmonotonic PDP networks actually provide an excellent example of the differences between symbolic and subsymbolic processing.

A Simple Artificial Neural Network



Value Unit



Michael R.W. Dawson

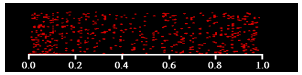
Much of our research involves finding ways to interpret trained networks that use this kind of unit

Wiretapping Value Units

We have found that wiretapping hidden units can lead to elegant and rich network interpretations

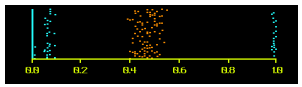
Jittered Density Plot

- One plot per hidden unit
- One point per pattern
- Horizontal location = activity
- Random vertical location prevents overlapping points



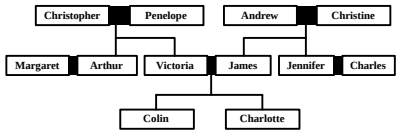
Banded Density Plots

- The jittered density plot for a value unit often reveals distinct, interpretable bands
- Patterns that fall in the same band share definite features



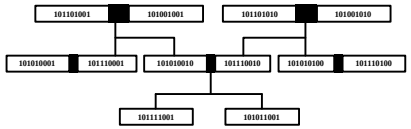
An Example Problem

- Hinton's kinship problem
- Ask a network about a name and a relation
- Network outputs a name
- "Who is James' father?" "Andrew"



Network Representation

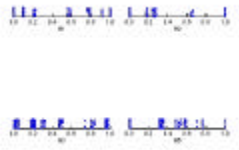
- 21 inputs, 6 hidden, 9 output
- 9 bit code for name
  (family, gender, generation, person)
- 12 bit unary code for relation
  (nephew, niece, aunt, uncle, brother, sister, father, mother, daughter, son, wife, husband)
- 6 families, 52 queries per family, 312 patterns



Family Detectors

In each of these units, every band represents a single family

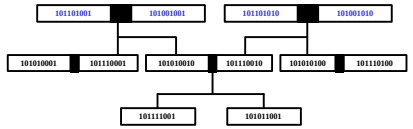
Tree Regularity Detectors



In each of these units, bands represent groups of individuals within a family tree

Example Band

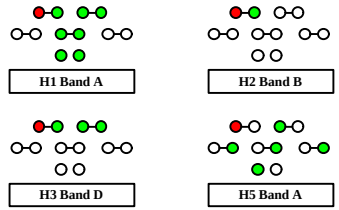
Hidden Unit 3, Band D, N = 24
wife or husband of person 010 in generation 1,
or
father or mother of person 010 in generation 2



Coarse Coding

- How are these broad categories of individuals used by the network?
- Individuals are represented by coarse coding
- One person falls out of the intersection of different bands in different hidden units

Example Intersection



Conclusions

- Some PDP networks can be interpreted
- Jittered density plots can be used to identify regularities in the hidden units of value unit networks
- Local features associated with bands in these density plots can be used to determine how a network solves a pattern recognition problem
- Course coding of features across hidden units can also be used to solve pattern recognition problems

References

Berkeley, I. S. N. (1997). What the #$*%! is a subsymbol?. Paper presented at the 1997 meeting of the Society For Exact Philosophy: Web version available at http://www.ucs.usl.edu /~isb9112/dept/phil341/subsymbol/subsymbol.html.

Berkeley, I. S. N., Dawson, M. R. W., Medler, D. A., Schopflocher, D. P., & Hornsby, L.(1995). Density plots of hidden value unit activations reveal interpretable bands. *Connection Science, 7*, 167-186.

Dawson, M. R. W. (1998). *Understanding Cognitive Science*. Oxford, UK: Blackwell.

Dawson, M. R. W., Medler, D. A., & Berkeley, I. S. N. (1997). PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology, 10*, 25-40.

Hinton, G. E. (1986). *Learning distributed representations of concepts*. Paper presented at The 8th Annual Meeting of the Cognitive Science Society, Ann Arbor, MI.

McCaughan, D. B. (1997, June 9-12). *On the properties of periodic perceptrons*. Paper presented at the IEEE/INNS International Conference on Neural Networks (ICNN'97), Houston, TX.