Extreme Attraction: On the Discrete Representation Preference of Attractor Networks

David C. Noelle (DNOELLE@CS.UCSD.EDU)
Garrison W. Cottrell (GARY@CS.UCSD.EDU)
Department of Computer Science & Engineering
University of California, San Diego
La Jolla, CA 92093-0114 USA

Fred R. Wilms (FREDLOCK@CS.UCSB.EDU)
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106-5110 USA

Overview

Attractor networks play an important role in many connectionist cognitive models. The dynamic behavior of these recurrent networks, which typically involves the convergence of activation levels to stable fixed-points over time, provides a means for modeling the time course of cognitive processes. Networks of this kind are often used as associative memories and as resolvers of soft constraint satisfaction problems.

Early attractor networks used step activation functions, allowing only two activation values for each processing element. Contemporary networks, however, often use sigmoidal activation functions. Continuous outputs allow for continuous target patterns, and these have at least two apparent advantages: (1) they allow for richer representational schemes, such as mapping activations to probabilities, and (2) they allow for a larger number of attractors. Attractor network models, like those for lexical semantics (Clouse and Cottrell, 1995) and consciousness (Mathis and Mozer, 1995), would seem to benefit from continuous patterns, and other models using continuous vectors, like the CHARM memory model (Metcalfe Eich, 1982), would benefit from attractor dynamics.

In this work we argue that these apparent advantages of continuous targets are illusory, at least when attractors are being learned using standard methods. Our simulation results indicate that attractor networks, even those with continuous activation functions, are best suited for use with target vectors consisting of polarized discrete elements.

Simulation Results

For this investigation we used single layer networks, with complete interconnections between processing elements (including self-connections), asymmetric weights, and sigmoidal activation functions. Unit activity spanned between -1 and $+1$. The networks were trained to form fixed-point attractors for a collection of target vectors using backpropagation-through-time, backpropagating the error signal for 10 time steps. A low learning rate of 0.001 was used, with no momentum. Networks were presented with each target pattern 10,000 times. We examined target patterns sampled randomly from the extreme corners of activation space (polarized vectors), patterns sampled uniformly over the whole activation space, and patterns sampled from the surface of the hypersphere inscribed by the whole activation space. At regular intervals during training, the dynamics of the networks were checked in regions near target patterns. The number of stable fixed-point attractors in these regions were recorded.

Rigorous experiments on small networks, ranging in size from 2 units to 16 units, revealed that the most attractors were learned when targets were placed in the extreme corners of activation space. An example of this result, for size 16 networks, is shown in Figure 1. Furthermore, attractors were learned *faster* (i.e., after fewer presentations) with corner targets. Lastly, corner target patterns resulted in attractors which were *closer*, in Euclidean distance, to their corresponding targets as compared to the other target distributions.

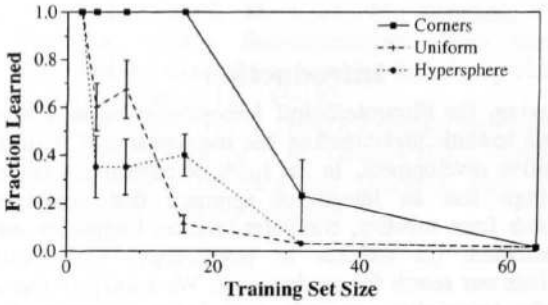

Figure 1: Fraction of the training set of target attractors actually learned as a function of the training set size.

These results were also found for larger networks, with 100 processing elements, trained for shorter durations, and also for sparse target patterns.

In short, attractor networks with sigmoidal units show higher capacity, faster learning, and greater accuracy when targets are placed in the extreme corners of activation space.

References

Clouse, D. S. and Cottrell, G. W. (1995). Lexical access with internet semantics. Presented at *Using High-dimensional Semantic Spaces Derived from Large Text Corpora* Symposium, *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pages 13–14.

Mathis, D. W. and Mozer, M. C. (1995). On the computational utility of consciousness. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances In Neural Information Processing Systems 7*, pages 11–18, Denver. MIT Press.

Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review*, 89(6):627–661.

Noelle, D. C., Cottrell, G. W., and Wilms, F. R. (1997). Extreme attraction: The benefits of corner attractors. Technical Report CS97-536, Department of Computer Science & Engineering, UCSD.