Learned Categorical Perception Effects in Neural Networks

Janet K. Andrews (andrewsj@vassar.edu)
Kenneth R. Livingston (livingst@vassar.edu)
Dalindyebo Shabalala (dashabalala@vassar.edu)
Vassar College Program in Cognitive Science
124 Raymond Avenue, Poughkeepsie, NY 12604 USA

Background

A number of recent studies have demonstrated what have come to be known as "learned categorical perception" effects in humans (e.g., Goldstone, 1994). The most common form of these effects is called "compression": People have greater difficulty discriminating or judge to be more similar patterns that they have learned to classify together, compared to people who have not learned to classify the patterns.

Harnad, Hanson, & Lubin (1995) trained simple, three-layer nets to categorize inputs coded to correspond to lines of different lengths. The nets were first required to produce as output lines identical in length to the inputs, and then trained to sort the inputs into the categories "short," "medium," "long" while maintaining auto-association performance. Additional output units marked category membership. There were strong compression/expansion effects in hidden-unit activation space, as measured by comparing the interstimulus distances for auto-association alone with those for auto-association with categorization.

The purpose of the present study is to determine whether categorical perception-like compression/expansion effects would occur in three-layer networks trained using back propagation to categorize a different, more complex set of inputs which do generate compression effects in human category learners (see Livingston & Andrews, 1997).

Experiments and Results

Inputs to the nets were generated from stimuli used in Livingston and Andrews (1997): drawings of the genitalia of six male and six female day-old chicks for training and two additional drawings for transfer testing (see the left side of Figure 1; the leftmost picutre is of a male, the one to its immediate right a female). Each drawing was mapped to a 25 by 20 grid (4 pixels per region) and the mean intensity of each region was determined. These values were presented to the 500 input units of each net.

A net with 500 output units was first trained to auto-associate the twelve stimuli. To prevent it from memorizing individual input patterns while still achieving satisfactory learning, six hidden units were used. Twenty output units were then added, with ten corresponding to each of the categories "Male" and "Female." The net was trained to classify the twelve input patterns. This process was repeated for ten different randomly seeded nets.

The nets' category training and transfer performance was comparable to that reported in Livingston and Andrews' (1997). To test whether the nets displayed categorical perception effects, the hidden unit activation levels of each input pattern were determined for each auto-association net and its expanded category learning version. Interpreted as points in a multidimensional space, distances were calculated between every possible pair of items and mean distance was determined for male-male pairs, female-female pairs, and mixed pairs for each net. Surprisingly, no compression was observed in the categorizing nets relative to their autoassociation only versions, i.e., within category item pairs did not move closer together following learning.

To analyze the nature of the hidden unit representations we converted the final weights from the input units to each hidden unit of the categorizing nets into greyscale images. This was done for a net using only two output units (one for each category) and two hidden units (HUs) that did not undergo autoassociation in an effort to maximize the category-relevant information extracted by the hidden units. These images are shown on the right side of Figure 1; the one shown furthest to the right performed well by itself on females but not males, and the other one was the reverse.

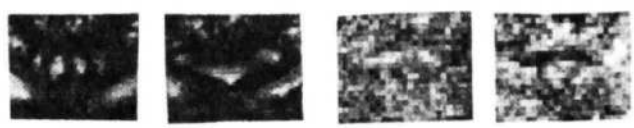

Figure 1: Sample stimuli (left) and HU images (right).

Acknowledgments

The research reported here was funded in part by the Vassar College Undergraduate Research Summer Institute. The order of authors is alphabetical.

References

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*, 178-200.

Harnad, S. Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps toward principled integration* . Boston: Academic Press.

Livingston, K., & Andrews, J. (1997). Prior theory effects on learned categorical perception. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 986). Hillsdale, NJ: Lawrence Erlbaum Associates.