Conceptual Hierarchies Arise from the Dynamics of Learning and Processing: Insights from a Flat Attractor Network

Christopher M. O'Connor (cmoconno@uwo.ca)
Ken McRae (mcrae@uwo.ca)
Department of Psychology, University of Western Ontario
London, Ontario, Canada, N6A 6C2

George S. Cree (gcree@utsc.utoronto.ca)
Department of Life Sciences, University of Toronto at Scarborough
Toronto, Ontario, Canada, M1C 1A4

The structure of people's conceptual knowledge of concrete nouns has traditionally been viewed as hierarchical (Collins & Quillian, 1969). This has occurred primarily because human performance in many tasks (e.g., typicality rating and superordinate to exemplar priming) appear to implicate hierarchies. Thus hierarchies have been transported, rather transparently, into theories of the structure of knowledge.

The goal of the present research is to show that behavior that, on the surface, appears to demand a hierarchical model can be simulated using a feature-based attractor network that contains a single level of semantics. In addition, this model provides insight into the role of similarity on the temporal dynamics of conceptual processing.

Network Architecture and Training

The network mapped from 30 wordform nodes (representing spelling/sound of a word) to 2349 semantic feature nodes (there were no taxonomic features such as <is a bird> in the network; these were used to establish 20 superordinate categories). Each feature unit corresponded to a semantic feature (e.g., <has wings>) from McRae et al.'s (2005) norms. The feature nodes were fully inter-connected with no self-connections. Thus, no hierarchy was implemented transparently in the model.

All 541 basic-level concepts were taken from the norms. Using continuous recurrent backpropagation, the network learned to map a 3-unit wordform for each basic-level concept to semantic features representing that concept. Superordinate concept learning was more complex because there were no unique target representations for them. On each superordinate learning trial, a wordform was paired with the representation of one of its exemplars. For example "fruit" was paired with the features of *apple* on one trial, *cherry* on another, etc. Crucially, exemplar representations were paired with the corresponding superordinate wordform equally often, so that typicality was not built into the training regime, and the network developed superordinate representations based on experience with the exemplars.

Typicality

Any model of this sort must account for typicality ratings. This was simulated by computing the meaning of a superordinate (e.g., *fruit*), computing the meaning of an exemplar (e.g., *cherry*), then calculating the similarity of the computed representations using a cosine similarity metric. The correlation between cosine and participants' rating of typicality was computed for each category. This correlation was high for most of the 20 superordinate categories, and predictions of typicality ratings were basically equivalent for the model and family resemblance.

Temporal Dynamics of Similarity

A number of experiments have shown roughly equal superordinate to exemplar priming (*fruit* priming *cherry*) for high, medium, and low typicality exemplars. Paradoxically, other studies show that basic-level concepts must be highly similar to support priming (and attractor networks simulate these effects). We conducted an experiment in which we found that the magnitude of priming was indeed similar for high, medium and low typicality items. Priming simulations showed the same result. Unlike features of basic-level concepts, superordinate features are partially activated from a wordform, rather than being activated essentially to 1 or 0 (due to a superordinate's mapping from one wordform to many featural representations). Because the feature activations are on the active part of the sigmoidal function, it is easy for a network to move from a superordinate representation to the representation of one of its exemplars, resulting in equivalent priming effects regardless of typicality. Thus, the model's temporal dynamics provide novel insight into these seemingly inconsistent results.

This research shows that a flat feature-based attractor network produces emergent behavior that accounts for human results that have previously been viewed as requiring a hierarchical representational structure.

Acknowledgments

This research was supported by NSERC OGP0155704 and NIH DC0418 and MH6051701 grants to Ken McRae.

References

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547-559.