UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Advances in Modeling Human Category Learning with DIVA

Permalink

<https://escholarship.org/uc/item/3m21f8nf>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 30(30)

ISSN

1069-7977

Author

Kurtz, Kenneth J.

Publication Date

2008

Peer reviewed

Advances in Modeling Human Category Learning with DIVA

Kenneth J. Kurtz (kkurtz@binghamton.edu)

Department of Psychology, Binghamton University
Binghamton, NY, USA 13902-6000

Keywords: category learning, classification, neural networks, autoencoder, computational models, selective attention, rules

Theoretical background

The DIVA (Divergent Autoencoder) model of human category learning (Kurtz, 2007) brings renewed vitality to a set of compelling explanatory principles that had fallen out of favor after failing to account for benchmark learning phenomena like the relative ease of acquisition of elemental category structures (Shepard, Hovland, & Jenkins, 1961). Core principles that distinguish DIVA from other error-driven adaptive network models like ALCOVE (Kruschke, 1992) and SUSTAIN (Love, Gureckis, & Medin, 2004) are: (1) a learning mechanism based on abstraction/recoding of inputs, as opposed to selective attention and association with fixed item representations; (2) representing knowledge about category instances in the connection weights, rather than in localist internal nodes of a network; (3) learning and classifying based on reconstructive success (goodness-of-fit) using an auto-associative mechanism that preserves, distorts, and infers features of the input in light of category knowledge, as opposed to computing the match between the input and reference points (exemplars, prototypes, rules).

The DIVA model incorporates these principles by treating categories as coordinated statistical models – each category is learned as a channel of an autoencoder network trained with standard backpropagation. The feedforward network consists of a set of nodes encoding the input features, a single set of hidden nodes for recoding all inputs, and distinct (divergent) sets of output nodes that generate a decoding or reconstruction of the input features in light of each possible category. The relative success of reconstruction on each channel determines the probability of classification. This learning process implements a modified form of principle components analysis (PCA) in which the recoding weights compactly encode the important variability in the training set and the decoding weights yield a construal of the input features in terms of each category. In sum, the system assesses how well an input accords with the statistical model underlying each category and interprets A/B classification as follows: To what extent is it possible to make sense of the available data as being the features of an A versus as being the features of a B?

A new design principle and new findings

This presentation focuses on a new design feature, as well as simulation results, that importantly extend the depth and breadth of the DIVA account. Specifically, an additional mechanism for generating classification responses based on the reconstructive outputs has been developed. Rather than assessing reconstructive success across all features, the use of *unidimensional evaluation* allows classification decisions to

be made based on the ability of each category channel to accurately predict a single feature. This attention-like mechanism is completely independent of the learning process, but it allows the model to exploit its fast mastery of diagnostic within-categories statistical properties. With this design feature, DIVA yields impressive fits to a range of category learning phenomena that were previously thought to depend on attentionally-mediated similarity to reference points or the use of hybrid/separate systems including an explicit, independent rule-learning component. Further discussion will address novel predictions and results extending DIVA to the domains of inference learning, unsupervised learning, classification of continuous-dimension stimuli, and structured representations in learning and comparison (e.g., Kurtz, 2005; Kurtz & Loewenstein, 2007; Levering & Kurtz, 2006).

References

- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Kurtz, K.J. (2005). On knowing the category before knowing the features. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Kurtz, K.J. (2005). Abstraction versus selective attention in classification learning. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Kurtz, K.J. (2005). Re-representation in comparison: Building an empirical case. *Journal of Experimental & Theoretical Artificial Intelligence*, 17, 447-459.
- Kurtz, K.J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, 14, 560-576.
- Kurtz, K.J., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory & Cognition*, 35, 334-341.
- Levering, K., & Kurtz, K.J. (2006). The influence of learning to distinguish categories on graded structure. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309-332.
- Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs*, 75 (13, Whole No. 517).