The Divergent Autoencoder (DIVA) Account of Human Category Learning

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, PO Box 6000
Binghamton University (State University of New York)
Binghamton, NY 13902 USA

Abstract

The DIVA network model is introduced based on the novel computational principle of divergent autoencoding. DIVA produces excellent fits to classic data sets from Shepard, Hovland & Jenkins (1961) and Medin & Schafffer (1978). DIVA is also resistant to catastrophic interference. Such results have not previously been demonstrated by a model that is not committed to both localist coding of exemplars (or exceptions) and the use of an explicit selective attention mechanism.

Introduction

The problem of supervised classification learning is of fundamental importance in both cognitive psychology and machine learning. Models of many kinds have been put forward offering powerful solutions. This paper presents a novel approach to supervised learning that shows considerable promise as an account of human category learning and as a technology for applied problems. The DIVergent Autoencoder (DIVA) network model takes as a starting point the back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986) and the reconstructive autoencoder architecture (McClelland & Rumelhart, 1986). Autoassociative systems are powerful learning devices that have been shown to implement principle component analysis and avoid local minima (Baldi & Hornik, 1989); to be extensible to non-linear function approximation (Japkowicz, Hanson, & Gluck, 2000); and to perform compression (e.g., DeMers & Cottrell, 1993). DIVA also draws on a design principle of multi-task learning mediated through a common hidden layer that been articulated in the ORACL model of concept formation (Kurtz, 1997; Kurtz & Smith, in preparation) as well as in the literature on neural computation (Caruana, 1995; Intrator & Edelman, 1997; Gluck & Myers, 1993).

Japkowicz (2001) developed an approach for applying unsupervised learning to binary classification that is close in spirit to the present proposal. An autoencoder is trained only on the positive instances of a category. Subsequently, inputs can be tested for membership in the category by evaluating the reconstructive success of the autoencoder. A new example that is consistent with the set of learned category examples will show minimal error while an example that is inconsistent will show a higher level of output error suggesting an inability to construe the input as a category member. Japkowicz (2001) demonstrated good results on binary classification problems by training a model to recognize examples of one class. Successful reconstructions are classified as members and rejections are assumed to belong to the other category. The approach is not extensible to n-way classification tasks or to cases where an A/B/neither classification response is required.

The innovation unique to DIVA is a method for converting any supervised learning problem into a form addressable by autoassociative learning. Traditionally, an autoassociative system is only capable of categorization to the extent that it picks out the statistical structure of a training set in a manner like clustering. This process suggests category formation in the sense that if a training environment is naturally organized in terms of sets of self-similar cases, the autoassociative learning system will extract that structure. Similar inputs are similarly represented and subsequent generalization behavior reflects these attractors. However, such a system has no capacity to acquire a classification scheme based on supervision that crosscuts the correlational structure of the training set.

The computational principle of divergent autoencoding offers an elegant solution to this problem using an autoassociative learning channel for each output class in an n-way classification problem. For a standard A/B classification learning task, one output channel is designated for reconstructing patterns labeled A by the teaching signal and the other is assigned to patterns labeled B. No output units are explicitly assigned to code for the categories themselves. The correct classification choice is used to select the channel on which to apply the targets (which are the same as the input). The architecture consists of an input layer, a shared hidden layer, and a set of autoassociative output banks. The pattern of connectivity is full and feedforward; all weight update is by back-propagation.

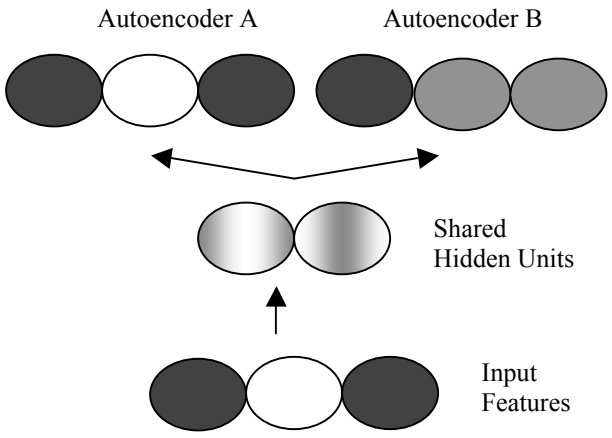

Figure 1: Architecture of the DIVA network.

The recoding of input information at the hidden layer of DIVA is shared by the set of channels, each of which is dedicated to learning to reconstruct the members of one class. This is different from forming compressed representations in traditional autoencoding and also differs from learning a recoding to achieve a linearly separable boundary between classes in a standard multi-layer perceptron architecture. DIVA will tend to produce different internal representations of an item depending upon the other same-category members included in the training set and also depending upon the contrasting categories being learned at the same time.

The DIVA network is tested by presenting an input which is processed along each channel in parallel. A classification response can be based on the amount of reconstructive error along a particular channel (i.e., testing the hypothesis that the example is a member of a particular category) as in Japkowicz (2001). In standard n-way classification tasks, the response is determined by selecting the class corresponding to the channel with the best reconstruction, i.e., the lowest sum-squared error. A version of Luce's (1963) choice rule is used to generate response probabilities for each choice K based on the inverse of the sum squared error at the output layer of the N channels. This is an extension of the common application of the choice rule to response generation based on output unit activations (e.g., Kruschke, 1992):

$$\Pr(K) = (1/SSE(K)) / \sum_{k=1}^{N} (1/SSE(k)), \qquad (1)$$

The logic of this paper is to demonstrate the power and promise of the DIVA network for cognitive simulation. To address the topic of human category learning, the primary goal is to evaluate the model on the two most widely studied datasets in the literature: the Shepard, Hovland, & Jenkins' (1961) dataset on ease of learning and the 5-4 learning problem introduced by Medin & Schaffer (1978). In addition to fitting benchmark data, a number of appreciable properties of DIVA in comparison with competing models will be outlined.

Experiment 1. The Relative Ease of Learning Across Category Structures

Shepard, Hovland, & Jenkins (1961) produced a groundbreaking analysis of the rate of acquisition of the six general types of category structures that are possible within a training set of binary-valued, overtly analyzable, three-dimensional stimuli. The most interpretable of these structures are: Type I, a unidimensional rule (UNI); Type II, the exclusive-or problem plus an irrelevant dimension (XOR); and Type IV, a family resemblance structure (FR). The results that have generated considerable challenges to model-builders is a qualitative ordering of the relative ease of learning: UNI fastest; followed by XOR; followed by roughly equivalent performance of FR, Type III, and Type V; followed by Type VI (though see Kurtz, under review).

The relative ease of learning the six types was tested across six random initializations of a (3-2-3x2) DIVA network (note: this refers to a DIVA network with three input units, two hidden units, and two autoassociative output channels with three units each). The number of epochs to criterion was determined based on total sum-squared error across the eight training patterns. Error was recorded only on the target-active (correct) channel. Two stopping points (SSE = .2; SSE = .1) were applied in accord with the strict criteria used in the behavioral study.

For the data reported in Table 1, learning rate of 0.25 and initial weight range of zero +/- 0.5 were used. However, qualitative performance was found to be consistent across variations in learning rate and the range of initial weight randomization. The only critical parameter is the number of hidden units. A simple systematic basis is used to determine the number of hidden units for a task. The smallest number of hidden units that can successfully reach asymptotic minimization of error across the manipulated learning conditions is the number that are used. This approach is in sharp contrast to the usual technique of exhaustive search through parameter space to find the best fit for each phenomenon of interest. In this case, two hidden units were required to consistently reduce error on the six SHJ types.

Table 1: Relative ease of category learning by DIVA

SHJ Type	Mean number of Epochs to criterion (0.2)	Mean number of Epochs to criterion (0.1)
I	566	840
II	847	1295
III	1195	1953
IV	1232	2087
V	1144	1750
VI	5719	9416

As can be seen in Table 1, the data are well fit (Type 1 < Type II < Types III, IV ,V < Type VI). Consistent findings were observed across the time course of training as was found in the SHJ replication by Nosofsky, Gluck, Palmeri, McKinley, & Glauthier (1994). By way of comparison, a standard feedforward (4-2-1) back-propagation network was tested under matching conditions. As also reported by Kruschke (1992), the network was far too quick to learn FR (comparable speed to UNI) and too slow to learn XOR. With two hidden units, some initializations became stuck in local minima (especially on Type V) and the system showed no progress on Type VI (a version of parity problem) without more hidden units.

Another way to test the performance of DIVA is to compute the classification response to each pattern using the choice rule over sum-squared error as outlined above. A single simulation was conducted in this fashion using the

identical set of initial weights for each of the SHJ types. The learning results after 500 training epochs appear in Table 2.

Table 2: Classification accuracy for a DIVA network.

SHJ Problem Type	Category Structure	Classification Accuracy
I	UNI	.97
II	XOR	.94
III		.84
IV	FR	.83
V		.93
VI		.56

The fit is excellent except for the overly good performance of Type V. There is a degree of variation across initializations of DIVA networks and in the case presented above, Type V showed performance at the upper end of its usual range. Such variation results primarily from the degree of consistency between the initial random configuration of weights and the form of the solution that is required. When lower learning rates and smaller initial weight variation are selected, the degree of variation lessens considerably.

In order to make clear how the learning occurs, DIVA solutions to the most interesting of the SHJ problem types (UNI, XOR, FR) are described as follows. A representative DIVA network solved the UNI problem (on F1) by assigning one hidden unit to code for the presence of F1 and F2 respectively. Each hidden unit strongly activated the F1 output units: via excitation on one channel and inhibition on the other. Each hidden unit also activated the appropriate non-diagnostic feature on each channel. F2 and F3 were always correct on the 'incorrect' category channel, while the output there for F1 was always exactly opposite to the input activation.

To solve XOR (on F1 and F2), a representative DIVA network largely ignored F2, but used signals from F1 and F3 to generate hidden layer recodings as shown in Table 3.

Table 3: Recodings formed by DIVA network on XOR.

Input	Hidden1 Activation	Hidden2 Activation	Target Category
101	0.2	0	0
001	0	0.8	1
000	0.8	1	0
011	0	0.9	0
111	0.2	0	1
100	1	0	0
110	1	0	1
010	0.8	1	0

The DIVA network used four areas of the activation space on H1 to code for the pairwise combinations of the diagnostic F1 and the non-diagnostic F3, while H2 primarily coded for F1. It is interesting to note that neither hidden unit explicitly represented the critical correlation between the diagnostic features (a standard back-propagation network would search for a recoding of the input specifically targeted to allow for linearly separable classification between the hidden layer to the output.) F1 and F3 were always correct on the 'incorrect' category channel, and the output there for F2 was always exactly opposite to the input activation.

On the FR problem, a representative DIVA network reached the following solution. H1 received an excitatory signal from F3 and an inhibitory signal from F2. H2 was sensitive to all three input features with a strong inhibitory signal from F1 and lesser excitation from F2 and F3 yielding the recodings shown in Table 4.

Table 4: Recodings formed by a DIVA network on FR.

Input	Hidden1 Activation	Hidden2 Activation	Target Category
101	1	0	1
001	1	0.9	0
000	0.4	0.6	0
011	0.5	1	1
111	0.5	0.4	1
100	0.5	0	0
110	0	0	1
010	0	1	0

The network assigned each input item to a unique location in the two-dimensional representational space of the hidden layer. The two channels showed equivalent connectivity projecting from the hidden layer and used strong bias weights to differentiate their performance. It is interesting to note that this solution parallels the behavior of an ordinary autoencoder operating on this training set. Once again, while operating entirely on the basis of the back-propagation algorithm, the hidden units do not act to transform the input for linearly separable classification. The 'incorrect' channel attempts to interpret each input as a member of its category and therefore produces markedly increased or reducing activation on one or more of the features.

The XOR problem holds a high place in the contemporary study of both human and machine learning. For decades, the connectionist tradition was halted by the lack of an algorithm to handle cases of *hard learning,* i.e., non-linearly separable functions. Rumelhart, Hinton, & Williams' (1986) paper on back-propagation of errors was a breakthrough that elicited tremendous productivity. The XOR problem remains a benchmark for evaluation of learning systems. A standard (hetero-associative) back-propagation network reaches asymptote on Type II learning (the XOR problem with an added irrelevant dimension) after approximately 3000 epochs of training. The DIVA network reached asymptote on average in 847 epochs. This nearly fourfold increase in speed of learning suggests that DIVA can perform non-linear function approximation with

considerable ease. The SHJ Type VI is the parity problem with three dimensions. The standard back-propagation network did not make any headway with two hidden units, but the DIVA network coasted smoothly down the error gradient. These findings suggest the power of the DIVA network as a general learning device.

In sum, the Shepard, Hovland, & Jenkins (1961) dataset is something of a litmus test for models of classification learning. Despite some question about the generality of the finding (see Kurtz, under review), it a seminal result in the literature. The design features of those models which have successfully fit this data have come to represent the state of the art in the field. Localist encoding and selective attention are core components of the three successful models: ALCOVE (Kruschke, 1992), SUSTAIN (Love, Medin, & Gureckis, 2004) and RULEX (Nosofsky, et al., 1994). These models all depend upon multiple free parameters (not including learning rate) that are selected according to the same data that is to be fit. RULEX uses three best-fitting parameters in addition to best-fitting attentional weights. ALCOVE and SUSTAIN each use three best-fitting parameters. DIVA offers a successful fit with a single parameter which is set a priori, rather than post-hoc, and offers a strong challenge to the widespread view that selective attention and localist representation are the correct explanatory constructs.

Experiment 2. Learning the 5-4 Categorization Problem

The case for the superiority of exemplar models has rested in no small part on extensive behavioral and computational tests of the 5-4 problem introduced by Medin & Schaffer (1978). A challenge has been raised recently (e.g., Smith & Minda, 2000) based on successful fits by a 'souped-up' version of a prototype model and questioning of the satisfactory nature of the exemplar account presented by Nosofsky, Kruschke, & McKinley (1992).

The 5-4 category problem consists of nine training items with four binary-valued features plus a set of transfer items. The design feature of the problem is that it is linearly separable (and therefore fair game for testing prototype models), but includes three very weak category members (for which only two out of the four features are consistent with the underlying prototype). Category B consists only of its prototype, one strong example, and two weak examples.

Model testing has focused not only on overall quantitative fit, but also to two qualitative aspects of the data. The first is that in non-elaborated experimental versions of the task, learners are more accurate on Stimulus A2 (which has two features in common with the A prototype) than they are on Stimulus A1 (which has three prototypical features). The prototype model predicts the opposite, while exemplar models capture the result (Nosofsky, et al., 1992). In addition, behavioral results typically show that a transfer test on the Category A prototype produces highly accurate responding, though not more so than the observed performance on training items that are somewhat distant from the prototype. Once again, the advantage goes to the exemplar view.

A (4-2-4x2) DIVA network was applied to the 5-4 problem using a learning rate of 0.1 and initial weights randomized in a range of zero +/- .05. The model was allowed to run for 1000 epochs. Performance on each training instance and the transfer items was determined by applying the choice rule to the sum-squared error along each channel. In terms of quantitative fit, a correlation of .96 was found between the probabilistic responses of the DIVA network and a summarization of thirty different behavioral tests of the 5-4 problem published by Smith & Minda (2000). The DIVA network produced a probability of A, Pr(A) = .96 for Stimulus A2 and Pr(A) = .85 for Stimulus A1; thereby fitting the critical qualitative result that was previously captured only by pure exemplar models and RULEX (Nosofsky, Palmeri, & McKinley, 1994). In addition, the transfer item T3 which is the prototype of Category A produced Pr(A) = .86 which was the strongest response to any transfer item, but was a lesser response than that shown for the training items A2 and A3. DIVA offers the first successful fit to these results by a model that does not implement the theoretical framework of localist encoding and selective attention.

Experiment 3. Avoiding Catastrophic Interference

Among some researchers, the phenomenon of catastrophic interference has been considered a fatal flaw for back-propagation as an account of human learning and memory (e.g., McCloskey & Cohen, 1989). In point of fact, a number of intriguing solutions and more nuanced treatments (McClelland, McNaughton, & O'Reilly, 1995; Mirman & Spivey, 2001) have appeared. Nonetheless, a minimal solution (one that does not graft an additional component, integrate additional mechanisms, or make modifications to the training set, etc.) has not been found. Is it possible to preserve the computational power and psychological validity of learning distributed internal representations via back-propagation without catastrophic interference?

The definitive demonstration of catastrophic interference for neural network models trained by back propagation is Ratcliffe's (1990) simulation result using the 4-4 encoder problem. The problem involves two learning phases. Training is performed to a certain level on the Phase I examples and then the training set is swapped. Phase II consists of training on *only* the second training set. The observed phenomenon is that the network performs well on the first training set at the end of Phase I, but the process of learning in Phase II "catastrophically" disrupts performance on Phase I examples. Phase I consists of three four-dimensional patterns to be autoassociatively reconstructed through an intermediate hidden layer. The patterns are: 1000, 0100, and 0010. Phase II consists of a single pattern: 0001.

Using DIVA, it is straightforward to assign a separate output channel to each sequential phase of learning. The

divergent autoencoding principle is applied in this case to separate phases of learning rather than to separate classification labels (as above). The same input and hidden units are used, however separate bank of outputs are used for each phase. Both channels are present in the architecture at all times, but targets only are applied to adjust the weights along the active channel. The critical assumption is that the shift between phases of learning must somehow be demarcated and psychologically encoded. The task context must make clear that "now you are to learn something else." In point of fact, traditional paradigms for studying interference usually make a very clear distinction between List 1 and List 2. An intriguing prediction is that an unannounced or non-obvious shift from Phase I to Phase II ought to elicit CI unless the switch is made manifest. As a final point of emphasis, no known model has been able to exploit the phase variable to prevent CI by devoting input or output units to code for the phase of each presented pattern.

A (4-3-4x2) DIVA network was tested with three hidden units and a learning rate of 0.2 in accord with Ratcliffe (1990) and Kruschke (1992). Weights were randomly initialized in a tighter range around zero. The network required 550 epochs to reach the 70% training criterion for Phase 1 learning used by previous investigators. As explained above, Phase I training applied targets only on the P1 channel. The same amount of training was conducted for Phase II on just the 0001 pattern using only the targets on the P2 channel.

Table 5: Output Activations of DIVA network on Sequential Learning Task.

Input	Channel for P1	Channel for P2
After Phase 1		
1000	.74 .19 .17 .04	
0100	.18 .68 .23 .03	
0010	.16 .24 .70 .04	
0001		.49 .49 .49 .50
After Phase 2		
1000	.73 .21 .15 .03	
0100	.16 .66 .24 .03	
0010	.17 .22 .68 .03	
0001		.04 .04 .04 .96

As shown in Table 5, catastrophic forgetting was fully avoided. Similar performance was observed across differently initialized runs and variations in learning rate and initial weight range. Two follow-up tests were conducted. The DIVA network was tested using negative valued (-1) input activations rather than zero-valued 'off' units. This also yielded successful results. In addition, an alternate version of Phase II learning was conducted using the pattern <–1 1 –1 1>. This extends the problem beyond the case in which positive activation of the features is segregated between the two training phases. Once again, performance on Phase I examples remained intact.

The success of the DIVA network can be explained very simply. The weights from Features 1-3 to the hidden layer are hardly affected by Phase II training, and the weights from the hidden layer to the P1 channel are affected not at all. However, this is not at all equivalent to using entirely different networks for the two phases of learning. The same input units, hidden units, and connecting weights are used. The two learning phases are equivalent for DIVA to learning a two-way classification problem with massed practice. One can interpret the DIVA solution to the problem of catastrophic interference as the establishment of a contextually-driven classification of inputs as members of either Phase 1 or Phase 2. With this one very plausible assumption, divergent autoencoding preserves the back-propagation machinery for error-driven learning without the catastrophic interference.

General Discussion

Given the demonstrated promise of DIVA, a number of further explorations are underway. DIVA shows a tendency to shift during learning from more general to more specific category representations (e.g., Smith & Minda, 1998). DIVA is naturally extensible to the recently vigorous investigation of category learning beyond traditional classification, i.e., inference learning, category use, unsupervised learning, and cross-classification. Since autoassociative processing naturally generates a feature-based representation as its output, applications to recognition memory, memory distortions, and feature prediction are forthcoming.

An intriguing aspect of the DIVA architecture is that it offers a straightforward mechanism for producing a convolved representation of any input in terms of any category known to the network. Imagine that a pattern representing a cat is presented to a DIVA network trained on various animal concepts. Regardless of which animal is the actual classification response, every channel produces an interpretation or construal of the input in terms of its category. The psychological nature of such construals is of great interest. For example, the similarity of concepts A and B can be computed as the degree of reconstructive success a DIVA network achieves in processing a prototypical example of A along a channel trained on concept B. Typicality or graded structure of category members can be understood as the degree of reconstructive success in processing a member of category A through the channel for that category. Argument strength for category-based induction can be understood as the degree of reconstructive success in processing a representation of the conclusion category along the channel(s) of premise categories. The internal representation generated by inputting a representation or representative example of one concept to the channel of another concept is likely to produce a conceptual combination or metaphoric interpretation. If a parsimonious means can be found to represent structural information in a form submittable to a neural network, the potential deepens.

In sum, DIVA provides an uncompromisingly good fit to the two most influential data sets on human category learning and does so with the following characteristics:

1. Distributed representation rather than localist nodes for individual instances
2. No selective attention mechanism
3. No performance-optimized free parameters

Therefore, the success of this model calls into question widely held theoretical assumptions. The DIVA network offers the brain-style computational power of back-propagation and overcomes its shortcomings in simulating human learning. The computational design principle of divergent autoencoding deserves consideration as an explanatory construct underlying broad aspects of cognition.

Acknowledgments

With thanks to David E. Rumelhart. This project was partially supported by NIH award 1R03MH68412-1.

References

Baldi, P. and Hornik, K. (1989) Neural networks and principal components analysis: Learning from examples without local minima. *Neural Networks, 2*, 53-58.

Caruana, R. (1995). Learning many related tasks at the same time with backpropagation, *Advances in Neural Information Processing Systems* Vol. 7, pp. 657-664, Morgan Kaufmann, San Mateo, CA, 1995.

DeMers, D. & Cottrell, G. (1993). Nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems Vol. 5*, pp. 580-587, San Mateo, CA: Morgan Kaufmann.

Gluck, M. A. & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3, 491-516.

Intrator, N. and Edelman, S. (1997). Learning low-dimensional representations via the usage of multiple-class labels. *Network 8, 259-281.*

Japkowicz, N. (2001). Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning, 42,* 97-122.

Japkowicz, N., Hanson S.J., & Gluck, M.A. (2000). Nonlinear Autoassociation is not Equivalent to PCA. *Neural Computation, 12,* 531-545.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kurtz, K.J (1997). The influence of category learning on similarity. *Unpublished doctoral dissertation*.

Kurtz, K.J. (under review). Abstraction versus selective attention in classification learning.

Kurtz, K.J. & Smith, G. (in preparation). The ORACL model of concept formation and representation.

Love, B.C., Medin, D.L, & Gureckis, T.M (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111, 309-332.

Luce, R.D. (1963). Detection and recognition. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), Handbook of mathematical psychology (pp. 103-189). New York: Wiley.

McClelland, J.L., McNaughton, B.L. & O'Reilly, R.C. (1995). Why There are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review, 102,* 419-457.

McClelland, J.L. & Rumelhart, D.E. (1986). A Distributed Model of Memory. In Rumelhart, D. E., McClelland, J.L. (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol II.* Applications (pp.170-215). Cambridge MA: MIT Press.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol 24, pp. 109-165). New York: Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Mirman, D. and Spivey, M. (2001) Retroactive interference in neural networks and in humans: the effect of pattern-based learning. *Connection Science*, 13(3), 257-275.

Nosofsky, R.M., Gluck, M., Palmeri, T.J., McKinley, S.C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352-369.

Nosofsky, R., Kruschke, J., & McKinley, S. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 211-233.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. K. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,55-79.*

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285-308.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol 1. Foundations* (pp.318-362). Cambridge, MA: Bradford Books/MIT Press.

Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13, Whole No. 517).

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1411-1430.

Smith, J.D. & Minda, J.P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 3-27.