Selecting Knowledge for Category Learning

Evan Heit (E.Heit@warwick.ac.uk)
Lewis Bott (L.A.Bott@warwick.ac.uk)
Department of Psychology; University of Warwick
Coventry CV4 7AL United Kingdom

Abstract

We present a category learning experiment in which subjects faced the knowledge selection problem, i.e., they needed to use their observations to determine which prior knowledge would be useful for learning. The issue of putting prior knowledge into neural network models is reviewed, and we present a new model which addresses the knowledge selection problem. This model gives a good account of the experimental results.

Introduction

At first glance, categorization would seem to simplify our lives, because a large number of individual observations can be classed together to allow reasoning and communicating about them as a group. But it has been pointed out that categorization itself entails further complexities. Medin and Ross (1997) noted that just 10 objects can be partitioned into categories over 100,000 different ways. So in addressing one computational problem, the high number of unique events, we are led to another computational problem, the high number of possible partitions of events. As a solution to this problem, it has been proposed that, by necessity, category learning is not entirely data driven (e.g., Peirce, 1931-1935). That is, people do not consider all possible partitions of observations when forming a category representation. Instead, we in effect consider a subset of the possibilities, using background knowledge for guidance. Indeed, it has by now been well established empirically that background knowledge has robust effects on facilitating category learning (see Heit, 1997, for a review).

Unfortunately, this solution itself raises yet another problem, namely the problem of selecting prior knowledge. There are many possible sources of background knowledge that could be helpful in learning about a new category. For example, imagine visiting some university campus for the first time and trying to learn about the general layout and architectural styles. Many sources of past knowledge could possibly be helpful, such as memories of other campuses or towns. In fact, it would be easy for the number of past observations to greatly outnumber the number of new observations! In light of this knowledge selection problem, how could background knowledge actually make concept learning easier?

The knowledge selection problem does seem very troublesome for experimental and computational approaches to category learning, but it is important to note that people do manage to solve this problem every day. In addition, it is encouraging to pick up any textbook on Bayesian statistics and find many techniques for combining multiple prior beliefs with observations, and selecting among these beliefs based on the data observed. In Bayesian statistics there is no assumption that a learner starts with optimal or perfectly correct prior beliefs. Instead, the learner begins with a reasonable guess that merely serves as an initial basis for learning, with corrective information then provided by the data. Indeed, it is possible to start with a whole set of different prior beliefs, with a distribution of initial degrees of confidence in each of these. When observations are made, confidence in various prior beliefs can be increased or decreased as appropriate. (See also Heit, 1998.) That is, observations can be used to select from a set of prior hypotheses.

Many previous experiments on knowledge effects on category learning have avoided the knowledge selection problem by more or less telling the subjects which prior knowledge to use. One of the exceptions is a study by Murphy and Allopenna (1994), in which subjects learned about categories of buildings, animals, and vehicles, with labels such as "Category 1" and "Category 2." These category labels did not constrain the knowledge selection problem very much. When a subject learned about a new category of vehicles, for example, there were many known types of vehicles that could be informative. It was impossible to know in advance whether to use prior knowledge about snowmobiles, ice cream vans, heavy trucks, or jeeps. However, the content of the category itself, that is, the descriptions of category members, were helpful in finding useful prior knowledge. For example, when subjects observed a category member with the description "made in Africa, lightly insulated, and drives in jungles," they were able to access knowledge about vehicles used in hot weather such as jeeps, rather than knowledge about other vehicles such as snowmobiles and heavy trucks.

Our own experiment was an attempt to further address the phenomenon of knowledge selection. Like Murphy and Allopenna, we used building categories. (Also see Heit and Bott, 1999, for an experiment with vehicle categories.) Given that people already know about many kinds of buildings, we see these stimuli as encouraging knowledge selection processes. Unlike Murphy and Allopenna, we collected data over the course of learning. One of our goals was to show that in some situations, categorization judgments are not affected early on by prior knowledge, until many observations have been made and relevant prior knowledge can be assembled. Therefore it was necessary to collect categorization judgments after various numbers of category members had been observed. Our general was that in terms of various measures there would be increasing knowledge effects over the course of learning. Another advantage of collecting data along the course of learning was that our data

were suitable for developing and testing a computational model of category learning.

We next present an experiment on knowledge selection in category learning, followed by a brief review of computational models that employ prior knowledge and then by the introduction of a computational model that addresses knowledge selection

Method

The 77 subjects learned about two categories of buildings, referred to as Doe buildings and Lee buildings. The subjects were told to imagine that they were reading a book with a series of descriptions of buildings. The stimuli were organized as five blocks, with descriptions of four Doe buildings and four Lee buildings presented in each block. Each description included the category label (Doe or Lee) and a list of featural information. There were two critical features presented in each description and two filler features. The critical features for each category were related to a known type of building (e.g., churches for Doe and office blocks for Lee or vice versa). In contrast, the filler features were general characteristics that could be true of just about any building. Finally, each description contained three pieces of individuating information (name of builder, surveyor, and photographer). The main prediction was that there would be increasing facilitation on critical features over the course of learning, as subjects were increasingly able to select useful prior knowledge.

The critical and filler features were derived from a pretest, which involved a series of sorting tasks in which subjects were asked to place each feature into one of two groups. After a series of iterations, replacing features as necessary, a set of 8 pairs of critical features and 8 pairs of filler features was obtained. A final pre-test group of 20 subjects sorted each of the critical features with at least 90% preferring one group over the other, and for the filler features preference for one group was always less than 75%. In addition, subjects were readily able to describe one sorted pile of features as being related to churches or old buildings, and the other as being related to office buildings or other commercial buildings. The complete list of critical features, as well as sample filler features, are shown in Table 1.

From the 8 pairs of critical features, 4 pairs were randomly assigned to presentation frequency one. Each feature in each pair was presented in one description per block, either Doe or Lee. Two pairs were assigned to presentation frequency two, and each feature presented in two descriptions per block. Finally, 2 pairs of features were not presented at all in the study blocks (but they were tested in test blocks). Likewise, the 8 filler features were assigned to presentation frequencies one, two, and zero.

There was a sequence of 5 study-test blocks. In each study block, the building descriptions, each with a category label, were presented individually, for 6 s each. A sample description would be: {Lee building type, Builder: T Jones, near a river, has gas central heating, Surveyor: R Rawson, Photographer: A Ferraro, has steeply angled roof, has wooden furniture}. Subjects were given memorization instructions. Following each study block was a test block, in which subjects were asked to categorize 40 single features, in the Doe or Lee categories. These test items included 24 individuating features, 8 critical features (4 presented once, 2 presented twice, and 2 not presented), and 8 filler features (same distribution as critical features).

Table 1. Critical and filler features for building stimuli.

Critical Features
has steeply angled roof, has a flat roof
has wooden furniture, has metal furniture
has an interesting structure, has a repetitive structure
old building, new building
quiet building, busy building
lit by candles, lit by fluorescent light
ornately decorated, blandly decorated
built with stone, built with metal and concrete
Sample Filler Features
near a bus station, not near a bus station
designed by a local architect,
designed by an international architect
has gas central heating, has electric central heating

Results and Discussion

Initial analyses did not reveal any significant differences between presentation frequency 1 and presentation frequency 2; therefore the results were pooled over these two presentation frequencies. The average proportions correct are shown in Figure 1. The top panel shows responses to features that had been presented during the study blocks. Overall, there is a trend for performance to improve over blocks. Although there is no difference between critical and filler features in the first block, the difference between the two kinds of features, that is, the gap between lines, widens after the first block, suggesting increased facilitation on critical features over the course of learning. The bottom panel shows responses to the features that had not been presented at all. Responses to filler features essentially represent chance responding. The responses to critical, non-presented features are more interesting. Even though these features were never presented in study blocks, categorization performance clearly improved from the first block to the fifth block, suggesting an increasing influence of knowledge.

The results were analyzed with a three-way ANOVA with block, feature type (critical or filler), and presentation (observed or not observed) entered as variables. Each of the variables had statistically significant main effects, and likewise each of the two-way interactions were significant. Perhaps the most important interaction was the feature type by block interaction, supporting the observation that the difference between critical and filler features increased across blocks.

Finally, performance on the individuating features increased steadily from 51% correct in block 1 to 59% in block 5, suggesting that subjects were devoting increased resources to learning the names on later blocks, as the other features were better learned.

The key result in this experiment was that subjects were increasingly influenced by background knowledge over the course of learning. One source of evidence for increasing influences of knowledge is the results for presented features. There was no difference in classification accuracy for critical and filler features after the first training block, but by the end of the second block subjects had apparently retrieved prior knowledge that facilitated performance on critical features compared to filler features. Realizing that the Doe buildings are church-like and the Lee buildings are like office buildings, for example, would help answer questions about critical features but not filler features. Although performance on critical and filler features continued to improve over the course of learning, the advantage for critical features was persistent. The other source of evidence for changes in knowledge effects is the judgments on non-presented critical features. Subjects were never told the correct category for these features during training blocks. The only way to classify these features correctly was on the basis of general knowledge about buildings. Performance on non-presented critical features improved over the course of learning, suggesting that subjects were increasingly relying on appropriate knowledge for making judgments about these features.

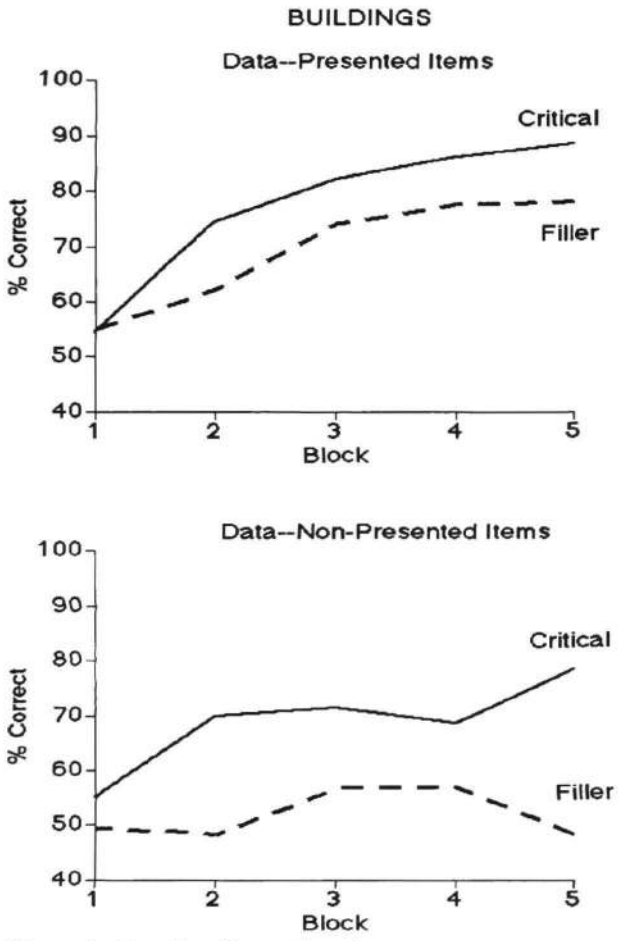

Figure 1. Results of experiment.

One surprising result was the lack of difference between features presented once per block and features presented twice per block. For both critical and filler features, we did not find any statistically significant difference in judgments for the two levels of presentation. It is tempting to relate this finding to Murphy and Allopenna (1994), who also found low sensitivity to frequency. Informal debriefing of subjects suggested to us that because each description, containing eight pieces of information, only appeared for 6 s, there may have been some strategic scanning of information. For example, in each block some subjects might have looked for features that had not already been presented in that block, hence overlooking a second presentation.

Putting Knowledge into Neural Networks

Next we set out to develop and apply a computational model that could address knowledge selection. We chose to work within the framework of neural network or connectionist models because they provide such a rich descriptive framework. That is, the complexity of connectionist models provides many opportunities for describing distinctive effects of knowledge on learning, as well as an appropriate framework for describing the dynamics of learning. Also, there has already been a great deal of research on different ways of putting knowledge into neural networks. Before we present our own model, we review some of this past work.

A useful framework for discussing prior knowledge in neural networks has been developed by Geman, Bienenstock and Dourstat (1992), who demonstrated that the generalization error when learning a concept can be broken down into a bias component and a variance component. Models that rely heavily on prior assumptions about the data, e.g., having architectural constraints that favor a particular conceptual structure, can lead to a high bias component, that is the model can persistently fail to capture aspects of the target concept which do not meet its prior assumptions. On the other hand, models that do not make strong assumptions about the concept to be learned can show a high variance component, that is they will be easily swayed by noise in training samples. Therefore a model without many assumptions could require an excessively large training sample to achieve good generalization. Further, reducing one type of error frequently is accompanied by an increase in the other type of error, leading to what Geman at al. referred to as the bias-variance dilemma. To reduce generalization error, both bias and variance must be reduced. We next review a number of learning algorithms that are aimed at reducing generalization error, keeping in mind the need to minimize the number of training examples as well.

One method for reducing the number of examples required for good generalization is to introduce "hints" into neural networks (e.g., Abu-Mostafa, 1995). Hints are general properties of a class of target concepts, independent of the specific details of the training data. Hints are introduced into the network by presenting "virtual examples" of the hint, and altering the error function to incorporate a term for the hint. Another approach to prior knowledge is to insert biases directly into neural networks by artificially setting the weights before learning begins. This approach has been

taken by, for example, Giles and Omlin (1993), whose method was to insert transition rules into recurrent neural networks that learned artificial grammars. Known transitions were built into the network and then unknown transitions were learned from the data.

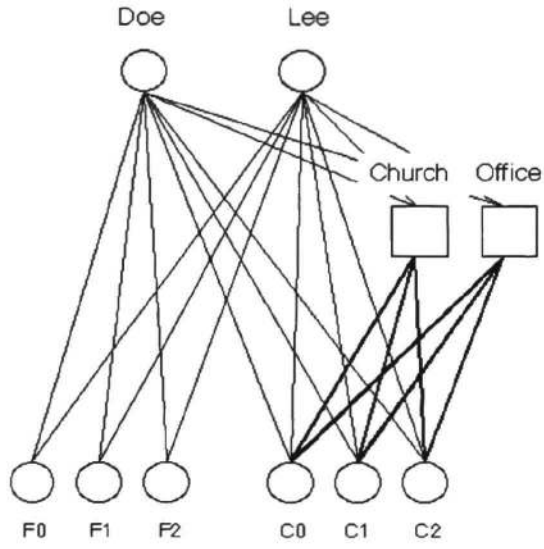

Figure 2. The Baywatch model.

Another way to build in prior knowledge is by varying the network architecture, to allow the network to have sufficient representational power to capture the underlying concept, but also avoid fitting the noise in the data. This goal is another way of looking at the bias-variance dilemma--a network that is too small leads to a high bias, but a network that is too large leads to high variance (and fitting the noise). Constructive networks (e.g., Prechelt, 1997) expand their architecture during learning, allowing the complexity of the network to increase as the data suggests it. Destructive networks, on the other hand, start off with an excess of hidden units and then prune off the hidden units which are not useful (e.g., Mozer & Smolensky 1989).

Rather than varying the network architecture over the course of learning, a different approach is to employ more than one architecture within a mixed network, and allow the network itself to learn which of the architectures is best for a particular problem. An example of this approach is the mixture-of-experts network (e.g., Jacobs, Jordan, & Barto, 1991). Jacobs et al. used a mixed network, with three modules having different structures (no hidden units, medium number of hidden units, and a high number). In effect, each module took a different approach to the bias-variance dilemma, with the simplest network being most constrained in terms of what it could learn and the network with many hidden units being most sensitive to variation in a training sample. The network was trained to perform two tasks, and it learned to allocate the module without hidden units to the simpler task while it allocated one of the modules with hidden units to the more complex task. We see the mixture-of-experts approach as coming close to the Bayesian idea of starting with multiple hypotheses then selecting among them based on the data.

The Baywatch Model

Our own approach to the knowledge selection problem has some parallels to the mixture-of-experts architecture, but instead of using modules with different structures, we used modules with different pools of pre-trained knowledge. Therefore our method also has some relations to techniques that insert prior knowledge directly into networks. Our model, illustrated in Figure 2, can be described as having one module or set of weights for strictly empirical learning. These weights do not get any pre-training. Then the model also has a set of experts which are pre-trained to recognize different known categories. For example, a network for learning about buildings might have experts which can recognize different kinds of buildings such as churches, office blocks, restaurants, and schools. (Only two of these expert modules are illustrated.) We refer to this model as the Baywatch model because it combines a general Bayesian approach to selecting among multiple sources of prior knowledge with an empirical learning component.

The Baywatch model is a feedforward network where the input units represent the individual features and the output units represent the Doe and Lee category nodes. The two hidden units correspond to two expert modules, or prior knowledge category nodes (PK nodes). The input units on the left side of Figure 2 represent filler features, and the input on the right side represent the critical features. The difference between the two types of features is that the filler features are only connected to the output nodes, whereas the critical features are connected both directly to the output nodes and indirectly to the output nodes via the PK nodes. The connections between the critical features and the PK nodes have fixed, pre-learned weights, so that values of critical features of the stimuli that correspond to church features would activate the church PK node, and likewise critical features of the stimuli that correspond to offices would activate the office PK node. The PK nodes have threshold functions, so that if any church feature, say, steeply angled roof, is presented, then the church PK node will be activated. The activation from the PK node would then be propagated to the output units.

In contrast to the connection weights between the critical features and the PK nodes, the other weights in the network are learnable through gradient descent. Adjusting the weights from filler units and the critical units to the output units allows the features to be associated with the category nodes in the empirical learning module. Finally, there are adjustable weights between the PK nodes and the category nodes. These represent the subject's capacity to associate known categories, say churches and office blocks, with the new categories, Doe and Lee buildings. We see this part of the network as addressing (at least in part) the knowledge selection problem, because here the network is learning to select from already known categories and apply this knowledge to judgments about new categories. (See Heit and Bott, 1999, for further details of the model and simulations.)

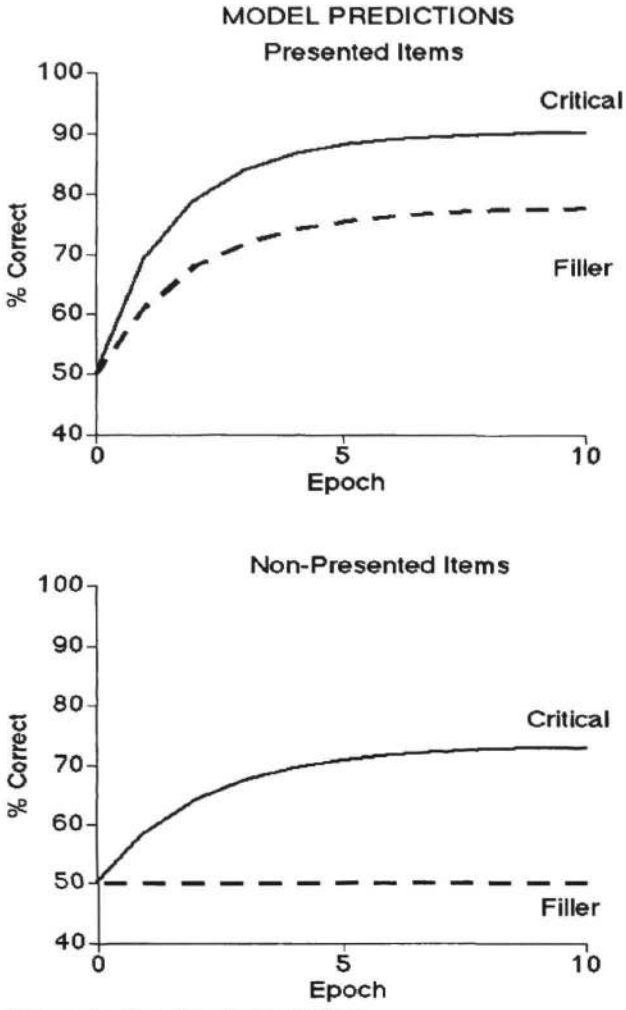

Figure 3. Results of simulations.

Simulations

The network was trained for a total of 10 blocks, with the learning rate in the delta rule set at 0.1 and the activation on each category node converted to a probability using a logistic transformation. The training stimuli consisted of four examples of buildings, two Doe exemplars and two Lee exemplars. Following each training block, the network was tested on the individual features by presenting a vector of all zeroes except for the particular feature of interest, which had a value of either +1 or -1. The results of the simulations are displayed in Figure 3. The top panel shows predictions for presented features, with the predictions for features presented once per block and features presented twice per block pooled together. The bottom panel shows predictions for features that had not been presented during training. The predictions fit well with the main results of the experiment. Critical features were learned more quickly than filler features, and critical features that hadn't been presented were responded to more accurately than chance, whereas filler features which hadn't been presented were at chance level.

To provide a better idea of how the Baywatch model uses prior knowledge, we re-ran the simulations without any PK nodes. In Figure 4, we show predictions on presented items, comparing versions of the model with and without prior knowledge. For critical features, in the top panel, it can be seen directly that prior knowledge does not have any influence initially on judgments; the model acts the same way with or without PK nodes. However, the beneficial effect of prior knowledge for critical features increases over the course of learning, as the network with PK nodes learns which categories to connect with its prior knowledge. In the bottom panel of Figure 4, there is evidence for a slight detrimental effect of prior knowledge on the learning of filler features. This result can be explained as a kind of overshadowing effect, in which knowledge of some highly predictive cues can reduce learning on other predictive cues.

One difference between the model's predictions and subjects' performance is that the model does predict more accurate judgments for features presented twice per block compared to features presented once per block. In contrast, there was no significant difference between these two levels of presentation in the experiments. This insensitivity to frequency could be an important aspect of concept learning in knowledge-rich domains but on the other hand it could just reflect subjects' reading strategies in this experiment. Therefore further experimental study is required.

Conclusion

How well would the Baywatch model scale up? The simulations were run with just two sources of prior knowledge (i.e., churches and office blocks) and the network was able to link up these two sources with the correct output categories, Doe and Lee. But people would obviously have a much larger number of known categories when facing the knowledge selection problem, due to large numbers of known kinds of buildings. In general, we think the model might scale up well, in terms of adding more prior knowledge nodes. It is useful to distinguish three different classes of PK nodes that might be added to the network in Figure 2, in addition to the church and office nodes.

First, irrelevant prior knowledge nodes might be added, which have little or no connection to the input stimuli. For example, there could be prior knowledge nodes for space stations, igloos, tents, and cave dwellings, added to the network, but these nodes would be hardly activated by the inputs. Therefore, adding PK nodes that are irrelevant to the stimuli would not affect the results of the simulations very much.

Second, additional PK nodes that are similar to the existing PK nodes might be included. For example, a PK node corresponding to cathedrals would entail much of the same connections to inputs as the church node. Putting in additional but similar PK nodes would enhance the prior knowledge effects but it would not really change their nature. Just as adding the PK node for churches helped performance on critical features of churches, relative to a straightforward empirical learning network (see Figure 4), adding another PK node for cathedrals would help even further. Paradoxically, there is no knowledge selection problem here, from adding another similar PK node. To the extent that sources of prior knowledge are mutually supporting, having multiple sources of prior knowledge need not harm performance.

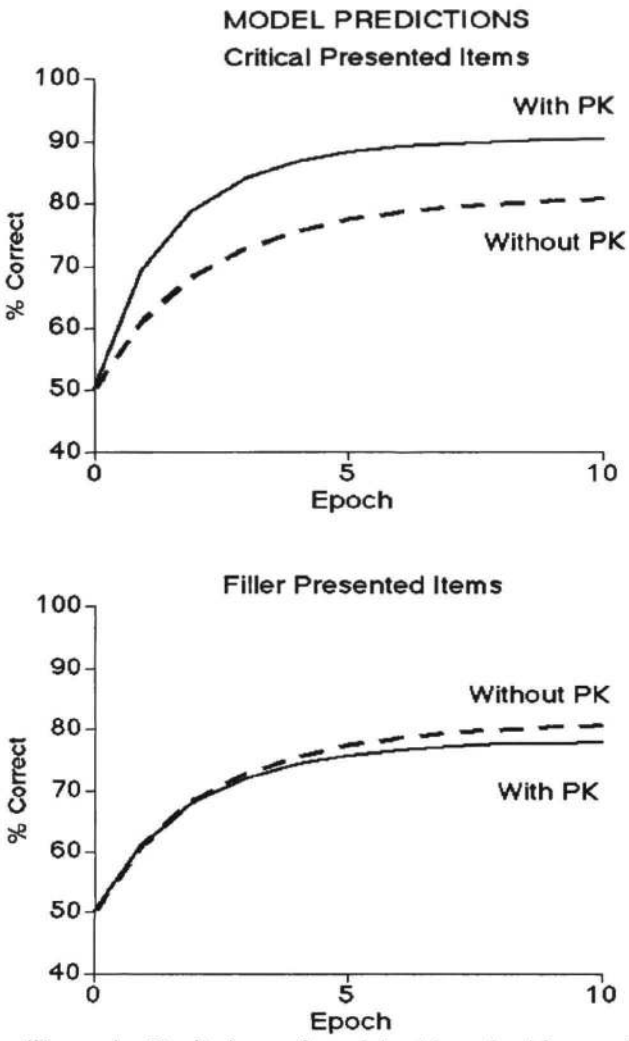

Figure 4. Predictions of model with and without prior knowledge.

Third, "malicious" prior knowledge nodes could be added to the network, for example, prior knowledge about some kind of building that is half-church and half-office block. Such PK nodes that are intermediate between the Doe and Lee categories might reduce the benefits of prior knowledge or even lead to costs due to knowledge, because they could make it more difficult to distinguish between the two categories.

More generally, we see the knowledge selection problem as having many facets. Certainly one of them is that when learning about novel categories, a learner would need to link up knowledge of familiar categories with judgments about the novel categories. The Baywatch model seems to address this aspect of knowledge selection, in terms of the gradual selection of prior knowledge nodes to use for a particular novel output category. In contrast, the prior knowledge in terms of connections from input units to PK nodes is fixed at the start of the simulations. It is assumed that these connections would have been already learned through ordinary associative processes, so that the network can more or less instantly recognize church or office buildings. However, there could be some gradual aspects of knowledge activation or retrieval that are not captured by the model. It could be the case that somehow the connections between input units and PK nodes would be learned over the course of making observations, so that the recognition of relevant categories in prior knowledge would not be instantaneous when a single observation is made. This aspect of knowledge selection might be studied more directly, for example by showing subjects a series of training examples and asking them to judge directly which familiar categories are related to these stimuli.

Acknowledgments

This research was supported by grants from ESRC and BBSRC.

References

Abu-Mostafa, Y. S. (1995). Hints. *Neural Computation, 7*, 639-671.

Geman, S., Bienenstock, E., & Dourstat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*, 1-58.

Giles, C. L., & Omlin, C. W. (1993). Extraction, insertion and refinement of symbolic rules in dynamically driven recurrent neural networks. *Connection Science, 5*, 307-337.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts, and categories* (pp. 7-41). London: Psychology Press.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition (pp. 248-274).* Oxford: Oxford University Press.

Heit, E., & Bott, L. (1999). Knowledge selection in category learning. In D. L. Medin (Ed.), *Psychology of Learning and Motivation.* San Diego: Academic Press.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture. *Cognitive Science, 15*, 219-250.

Medin, D. L., & Ross, B. H. (1997). *Cognitive Psychology.* (2nd ed.). Fort Worth: Harcourt Brace.

Mozer, M. C., & Smolensky, P. (1989). Using relevance to reduce network size automatically. *Connection Science, 1*, 3-16.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

Peirce, C. S. (1931-1935). *Collected papers of Charles Sanders Peirce.* Cambridge: Harvard University Press.

Prechelt, L. (1997). Investigation of the CasCor Family of Learning Algorithms. *Neural Networks, 10*, 885-896.