Cats could be dogs, but dogs could not be cats: what if they bark and mew? A Connectionist Account of Early Infant Memory and Categorization

Robert A.P. Reuter (rreuter@ulb.ac.be)
Cognitive Science Research Unit, CP 191, Av. F.D. Roosevelt, 50
B-1050 Brussels, Belgium

Abstract

The goal of this paper is to replicate and extend the connectionist model presented by Mareschal and French (1997) as an account of 'the particularities of […] infant memory and categorization'. With infants, the sequential presentation of cats followed by dogs yields an expected increase in infants' looking time, whereas the reversed presentation order does not. This intriguing asymmetry of infants' category formation, first reported by Quinn, Eimas, and Rosenkrantz (1993), was simulated by Mareschal et al.'s simple connectionist network. In addition, the authors proposed that this asymmetric categorization is a natural byproduct of the 'asymmetric overlaps of the visual feature distributions' of cats and dogs. Using a simple feedforward backpropagation network, we successfully replicated this asymmetric categorization effect, as well as a reported asymmetric exclusivity effect in the two categories, and an asymmetric interference effect of learning dogs on the memory for cats, but not of learning cats on the memory for dogs. We furthermore investigated the authors' explanation of the asymmetric effects, firstly, by systematically varying the overall similarity between learned items and interfering items, and secondly, by adding a binary feature to the input set, namely the animal cry (barking vs. mewing). The results of the present modeling underscore the authors' explanation of the observed effects in infants' memory and categorization, but also suggest lines of further experimental research susceptible to undermine the proposed connectionist account.

Introduction

In this paper, we report on a replication and two extensions of Mareschal and French's (1997) simple connectionist model that accounts fairly well for unexpected findings observed in spontaneous category formation in young infants (e.g., Quinn, Eimas, & Rosenkrantz, 1993) and in infant memory. Indeed, Mareschal et al. focused their modeling efforts on three target behaviors of very young infants concerning their categorization and memory, namely: (a) the ability to categorize complex visual stimuli, (b) the asymmetric effect in early categorization, and (c) interference effects in early memory.

The empirical data Mareschal et al. modeled show that infants, aged 3 to 4 months, are able to accurately categorize cats and dogs, and that they form an exclusive category of cats (thus excluding novel dogs) but an inclusive category of dogs (thus novel cats may well fall into the category of dogs) (Quinn et al., 1993). Furthermore, infants are known to present catastrophic forgetting of previously learned stimuli when shown certain other intervening material (e.g., Cohen & Gelber, 1975). This interference effect decreases precisely when infants' categorization abilities increase (Quinn & Eimas, 1996). Infant memory and categorization can thus be thought to be closely linked to each other and to depend on the same basic mechanisms.

Indeed, the connectionist account given for the observed asymmetric effects in categorization and for the interference effects in memory is based on the same reasoning. The asymmetry in category formation arises from the unequal overlap of the visual features distribution and not just the variance of the distribution itself. In other words, the values of the cat features fall within those of the dog values. Hence, based merely on the statistical structure of the input features, infants (and neural networks) form a category of dogs including cats, whereas they form a category of cats excluding (some) dogs[1]. The correlational structure extracting mechanism is precisely what accounts for the observed asymmetric effects of unequally exclusive (or inclusive) cat and dog categorizations. Catastrophic forgetting, on the other hand, can be understood as the deleterious effect of representing, within the same connections, stimuli whose features are very differently distributed. This, consequently, "washes out" the relevant knowledge previously stored in the network. However learning items whose features lie within the same range as those learned previously should not create very different internal representations - and may hence even consist in "learning more of the same". Based on this connectionist account of the relationship between infant categorization and memory, Mareschal et al. successfully produced the conjectured asymmetry in catastrophic interference consistent with the asymmetry observed in category elaboration (i.e., dogs interfered with previously learned cats, whereas cats did not interfere with learned dogs).

[1] Indeed, some dogs "look" like cats, since that they fall into the range of the cat distribution for some features. We will come back to this later.

In the following we, first of all, report on a conceptual replication of Mareschal et al.'s modeling results, concerning the development of categories, the asymmetry in category exclusivity, and the asymmetry in interference effects. Next, we show that the results of a cluster analysis performed on the input data and on the hidden units activation patterns (after learning of all items had occurred) suggest that the asymmetry in category exclusivity closely depends on the particular items used for "cross-category compatibility"[2] testing. Then, we explain how we tested Mareschal et al.'s connectionist account of asymmetry in category exclusivity and in interference effects, based on two systematic manipulations of the overlap in feature distributions of cat and dog categories. The first variation of overlap was produced by carefully choosing the items presented for training and those for interfering. The second variation of overlap was produced by the introduction of a supplementary (binary) input feature that unambiguously, taken individually, separates cats from dogs. Finally, we propose lines of further experimental research that might undermine the connectionist account embraced here.

The Model

We made the same assumptions as Mareschal et al. about the mapping between experimental results found in infants with the technique of preferential looking times and modeling results. The increase of error in the model's output is indeed assumed to be related to increased infant attention. The results reported below are based on the performance of a standard 10-8-10 feedforward backpropagation network, as well as that of a standard 11-8-11 network when the supplementary input feature was added. For both models, learning rate and momentum were both set to 0.9. The Fahlmann offset used in the original paper was not used in our model. Networks learned the data for a maximum of 250 epochs or until all output bits were within 0.2 of their targets. Results are averaged over 50 replications. Weights were updated after each stimulus presentation. Further details about changes in procedure compared to Mareschal et al.'s model are given in the results section below.

The Data

The data were identical to those used by Mareschal et al. A description of their origin and their characteristics can be found in their paper. In the following paragraphs, we report a cluster Analyses performed on the input data (and on the activation patterns of the trained network's hidden units) that will provide some insight into the inherent structure of the input data and its implication on modeling results. They will as well allow us to justify the proposed extensions of Mareschal et al.'s neural network simulations.

Cluster Analysis

On the basis of the clusters of correlated values of the input data, cats and dogs cannot be clearly separated into two mutually exclusive categories (see figure 1). Likewise, the connectionist model's internal representations (as reflected by the activation pattern of its hidden units after training on the whole set of stimuli) do not reflect clear-cut groupings of just dogs or just cats (see figure 2). Indeed, some cats, respectively some dogs, are more prototypical in terms of their overall feature similarity with the other members of their category.

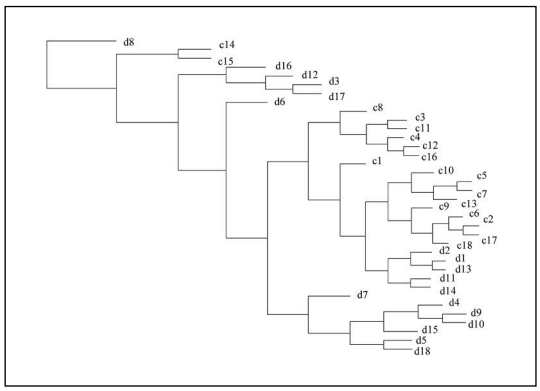

Figure 1. Cluster analysis performed on the input data patterns for cats (c1-c18) and dogs (d1-d18). Distances and cluster structure in this graph correspond to the overall similarity structure of the input patterns.

Moreover, and this is most crucial for Mareschal et al.'s proposed account of infant memory and categorization, nearly all cats (except 2 out of 18) fall into the "family" of "dog-cat"s[3], whereas, only 5 dogs (out of 18) clearly fall into the group which the majority of cats belong to. The clustering of cats and dogs into different subgroups suggests that the observed category exclusivity effects, as well as interference effects, should not occur for all combinations of learned and novel items (respectively, learned, interfering and novel/same-category items). Based on these results we predicted, precisely, that the closer items are in terms of cluster analysis distance, the smaller interference and cross-category exclusion effects should be.

[2] Cross-category compatibility refers to the question whether a novel exemplar is "accepted" as a member of the opposite category or not.

[3] "Dog-cat"s corresponds to a regrouping of clusters containing cats and dogs.

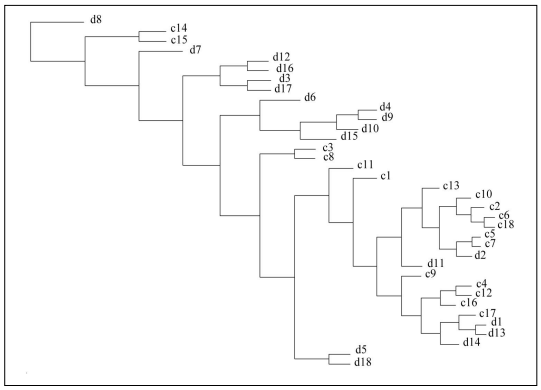

Figure 2. Cluster analysis performed on the activation patterns of the network's hidden units for cats (c1-c18) and dogs (d1-d18). Distances and cluster structure in this graph correspond to the overall similarity structure of the network's internal representations of the input patterns.

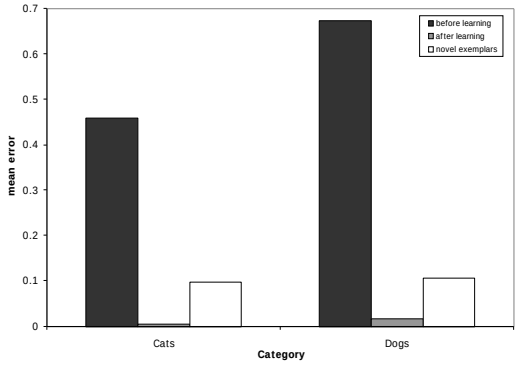

Figure 3. Mean error on the network's output when (a) presented with exemplars before learning, (b) presented with trained exemplars after learning, and (c) presented with untrained exemplars after learning.

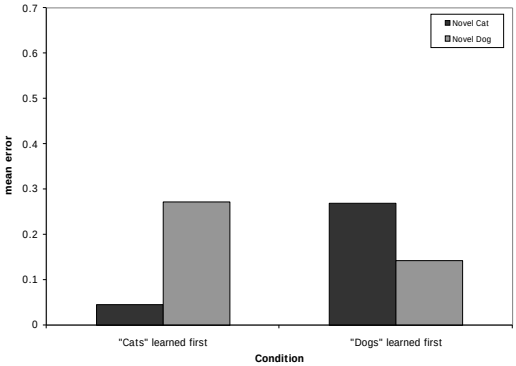

Figure 4. Asymmetric exclusivity of the cats and dogs categories. When trained first on cats, an untrained dog results in a larger increase of error than an untrained cat, but when trained on dogs, an untrained cat only produces a small increase in error as compared to an untrained dog.

Results

In the following section, we will briefly report the basic replication results since they nearly reflect the findings of Mareschal et al., and then describe our new results with their model.

The Development of Cat and Dog Categories

We obtained results comparable to those described in Mareschal et al. Networks do form a category of both cats and dogs. Figure 3 shows the initial mean error score, the mean error after training on the first 12 items of each category[4], and the mean error score (after learning) for the 6 remaining exemplars of the corresponding category (same-category testing). Needless to say that networks develop a faithful internal representation of both cats and dogs, and that they nevertheless recognize novel items as unfamiliar (small increase in error compared to the learned items). It is worth noting however that the initial error scores are slightly different between dogs and cats. Relative to those of cats, the features of dogs are more variable. Thus the mean error on dogs without training tends to be bigger than on cats. This confirms Mareschal et al.'s data analysis in terms of means and variances of the feature distributions of cats and dogs.

The Exclusivity of Cat and Dog Categories

Figure 4 shows the mean error on output of networks trained on 17 (out of 18) cats when they are presented with either the single remaining cat or with any of the dogs (18 out of 18), as well as with the corresponding opposite configuration (dogs learned first and then tested on novel dogs, respectively novel cats).

We used this training regime, instead of training the networks on 12 out of 18 items and then testing it on 4 cats, in order to get a reasonable number of different controlled combinations of training and testing items. All cats, respectively dogs, appear once as the novel item which the network has to categorize based on its nearly perfect "knowledge" of the same or opposite category. Our results show that, on average, dogs are less likely to be accepted as members of the cat category, than vice-versa. Indeed, a novel cat presented to a network that has learned dogs is less likely to produce a big increase in error, than a dog when presented to a network that was trained on cats. Based on the results of the cluster analysis, we suggest that the exclusion of the various cross-category items (but also of very atypical same-category items) somehow depends on their similarity with the core representation the network has developed during training.

[4] We suppose that this selection is pseudo-random, since we did not give the stimuli set a particular a priori order.

The Asymmetric Interference Effect

In this section we examine the effect of learning items of a second category on the networks ability to correctly "accept" novel items as belonging to the category it was trained on in the first time. The network, for instance, was trained on 12 cats, then tested on the remaining (novel) cats, then trained on 4 dogs, and finally re-tested on the novel cats. The difference in error scores on the novel cats before and after learning the 4 dogs is called interference effect, interference of dogs on the memory for cats. The same reasoning holds for the opposite case, memory for dogs interfered by training on cats. The interference effect of dogs on cats was easily replicated, by pseudo-randomly (see above) choosing cats to learn and dogs to interfere with the memory for cats. However, in contrast to Mareschal et al., we did observe a considerable interference effect of learning cats upon the memory for dogs.

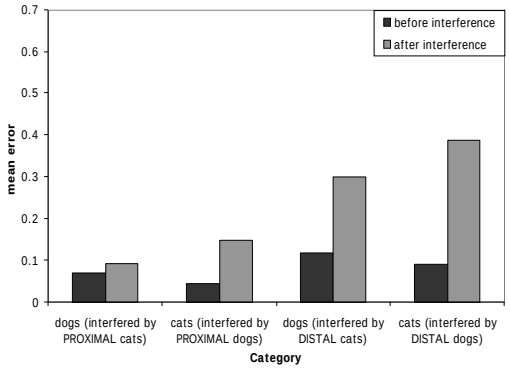

Figure 5. Network performance with untrained exemplars before and after learning an interfering category as a function of overall similarity between trained items and interfering items (distal vs. proximal).

See figure 5 for the results on the network's performance with novel items before and after learning items of the opposite category. We thus observed catastrophic interference in both cases, when not controlling for the similarity between training, testing and interfering. In figure 5, "distal" refers to items that are very dissimilar to the target category based on the cluster analysis. We thus analyzed the similarity between training and interfering items, we had pseudo-randomly chosen, and furthermore carefully selected packages of items in order to produce the "desired" results, based on the distance between items estimated by cluster analysis. The results of the simulations conducted on these items is reported in the following paragraph.

The Effect of Similarity between Learned and Interfering Items

We trained the network with various combinations of category learning items and interfering items from the other category. The choices of the items were motivated by the results of the cluster analysis performed on the input data. This enabled us to distinguish between groups of stimuli that are more or less "similar" to each other. We predicted that interference depends upon the overall feature similarity (i.e., overlap of feature distributions) between the central tendencies of these two groups, the bigger the similarity, the smaller the interference should be and vice-versa. We were, by this subterfuge, able to qualitatively reproduce Mareschal et al.'s asymmetry in interference effects by showing that, under selected conditions, dogs are not interfered by cats (see figure 5, dogs interfered by proximal cats). We were also able to show that results contrary to those reported in Mareschal et al. could be found. If cats learned in the first place, were "interfered" by very similar dogs, and the test items (i.e., novel cats) were chosen as close to the (two) learned set(s), then interference was minimal and relatively close to that observed in the case of dogs "not-interfered" by (similar) cats. Our results show that the results found by Mareschal et al. need not be the only possible ones. Furthermore, we could not manage to replicate the reported "average"[5] absence of interference of learning cats on the memory for dogs. Still, consistent with their connectionist account of infant category and memory our results are conclusive regarding the influence of overall feature similarity on the exclusivity and interference effect, though locally and not really globally. We could indeed show that overall similarity of items used for training and those used for interference could be used to predict interference of learning a second category on the memory for a first category. What we could not show was that this really is true for cats and dogs, on average, taken as categories. We remain agnostic to the very reasons of our failure to reproduce the absence of interference of cats on the memory for dogs when exemplars of both categories were randomly chosen. If the account given by Mareschal et al. is as general as they presume, a claim to which we subscribe in principle, then it should show up more consistently and more reliably with nearly any random selection of cats and dogs. Although we stay puzzled concerning the precise reasons of our failure to replicate, we have some hints on potential explanations. In fact, we wonder whether the rather small increase in error on dogs after interference by cats is consistent with the nevertheless not so small increase in error found with novel cats when dogs have been learned first. If it is true that, on average, learning cats does not have a deleterious influence on the memory for dogs, then why does presenting cat after training on dogs produce an increase in error compared to

[5] Since the authors did not precisely mention how they chose their items for training and interfering, we assume that the reported asymmetry in interference effects must be supposed to reflect average results, which is quite consistent with their account of the category exclusivity effects.

presenting a dog? After all, if learning cats is truly somewhat equivalent to learning more of the dog category[6], then why does testing on novel cats not produce less error then testing on novel dogs? We think that this prediction might be consistent with the connectionist account proposed by Mareschal et al., solely based on the asymmetry of overlap of feature distributions. Since nearly all cats belong to a narrow range of distribution embraced by the distribution of the dog category, dogs are more likely, on the average, to fall outside this narrow range, within which the "prototypical" dog also falls and which is precisely what the network's internal representation should reflect. Comparing cats and dogs to this "prototypical" dog should show that cats are situated closer to it in nearly all considered feature distributions, thus producing less error than dogs (which are more likely than cats to be different from this prototype).

The Effect of Induced Changes in the Inherent Structure of the Stimuli

We conducted the same simulations as before, except that networks were trained on inputs that included the animal cry (barking vs. mewing) as an eleventh characteristic variable. A cluster analysis (shown in figure 6) performed on these input data shows that this manipulation manifestly segregates dogs and cats into two next-to-perfectly distinct categories. The increase in distinctiveness of cats and dogs produced by the addition of the binary variable should eliminate (or at least considerably reduce), we hypothesized, the asymmetry of the category exclusivity, as well as the asymmetry of cross-category interference effects.

The results shown in figures 7 and 8 confirm our predictions based on the account given by Mareschal et al. In other words, cats no longer could be part of the dog category, and learning cats considerably interfered with the memory for the dog category. The inherent correlational structure of the input data thus has an important effect on networks categorization and memory. Note that the networks successfully formed a category of cats and dogs just like when the animal cry was not added to the stimuli features.

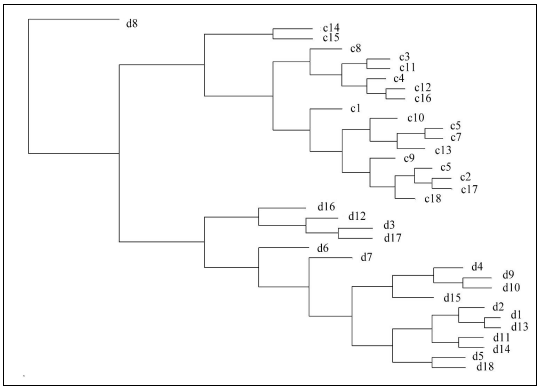

Figure 6. cluster analysis performed on the input data patterns for cats (c1-c18) and dogs (d1-d18) after addition of the “animal cry” feature. Distances and cluster structure in this graph correspond to the overall similarity structure of the modified input patterns.

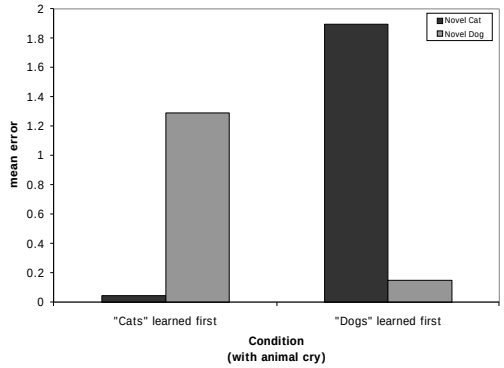

Figure 7. Symmetric exclusivity of the cats and dogs categories, when “animal cry” is added to the input features.

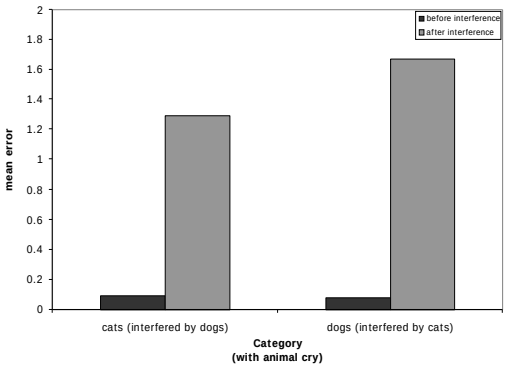

Figure 8. Network performance with untrained exemplars before and after learning an interfering category, when “animal cry” is added to the stimulus features.

[6] This is of course only true in some sense, since the dog category is clearly characterized by greater feature variance.

Discussion

Connectionist autoassociator networks, like young infants, form categorical representations of cats and dogs. The categories developed show asymmetric exclusivity, closely related to the unequal distribution of features in the stimuli shown. Most of the cats could be classified as dogs, but most dogs are not plausible cats. The model also suggests the presence of asymmetric interference effects of sequential learning of cats and dogs. Such effects have recently (Mareschal, French, & Quinn, draft) been observed in infants. Supported by empirical works on infants, the model thus pleads for a close link between the mechanisms underlying infant visual memory and categorization. In fact, Mareschal et al. (1997) claim that some kind of associative, data driven mechanism underlies early visual memory and categorization. The present paper underscores this claim, by showing that explicit manipulations of the correlational structure of the data input influences the networks performance. Connectionist models, by making clear assumptions about the input data, can thus be helpful in predicting which stimulus features are likely to be taken into account by infants. Indeed, experimental works (Spencer, Quinn, Johnson, & Karmiloff-Smith, 1997) have shown that infants typically rely 'on head/face information to categorically differentiate between cats and dogs, under conditions of short exposure duration'. This empirical result is consistent with the connectionist model presented here, since face visual features (viz. nose length and nose width) are the most informative features about the cat/dog distinction. In addition, based on Mareschal et al.'s (1997) connectionist account, it seems reasonable to predict that presenting pictures of cats and dogs in association with the corresponding animal cry should produce the same results in infants than in networks, namely mutual and symmetric category exclusivity and symmetric interference effects. Based on simulation results, we also predict a close parallelism between infants' and connectionist models' performances in memory and categorization task in terms of the similarity of particular stimuli (and combinations of stimuli). Truly, the model incorrectly excludes certain stimuli and not others, thus infants should present the same behavior pattern with precisely those stimuli in question. Likewise, if the model, for certain items but not others, does not present the discussed asymmetry in interference effects then infant's behavior should qualitatively reflect the same catastrophic forgetting. We thus suggest an item based analysis of networks' and infants' memory and categorization performances. Finally, we would like to recall that, by construction, connections networks' performance depends upon the very selection of certain stimulus features and not other. Thus, if networks produce categorization and memory effects similar to those of infants, then the selection of the particular features is given support. Nevertheless, it must be experimentally shown that infants actually rely on those features and not others. Still, connectionist models provide good predictions about which stimulus features are most likely to participate in infant categorization and memory.

Acknowledgments

Robert A.P. Reuter is a Research Assistant of the National Fund for Scientific Research (Belgium). Thanks to Bob French and Axel Cleeremans for their insightful comments on earlier drafts of this paper.

References

Cohen, L. & Gelber, E. R. (1975). Infant visual memory. In L. Cohen & Salapatek (Eds.), *Infant perception: From sensation to cognition*, Vol. 1 (pp. 347-403). NY: Academic Press.

Mareschal, D., & French, R. M. (1997). A Connectionist Account of Interference Effects in Early Infant Memory and Categorization, In *Proceedings of the 19th Annual Cognitive Science Society Conference*, NJ: LEA, pp. 484-489).

Mareschal, D., French, R. M., & Quinn, P. (draft, April 14, 1998). Interference Effects in Early Infant Visual Memory and Categorisation: A Connectionist Model.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. In C. Rovee-Collier & L. P. Lippsitt (Eds.), *Advances in infancy research* (Vol. 10, pp. 1-36). Norwood, NJ: Ablex.

Quinn, P. C., & Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants, *Perception*, 22, 463-475.

Spencer, J., Quinn, P. C., Johnson, M. H., & Karmiloff-Smith, A. (1997), Heads you win, tails you loose: Evidence for young infants categorizing mammals by head and facial attributes, *Early Development and Parenting*, Vol. 6 (3-4), 113-126.